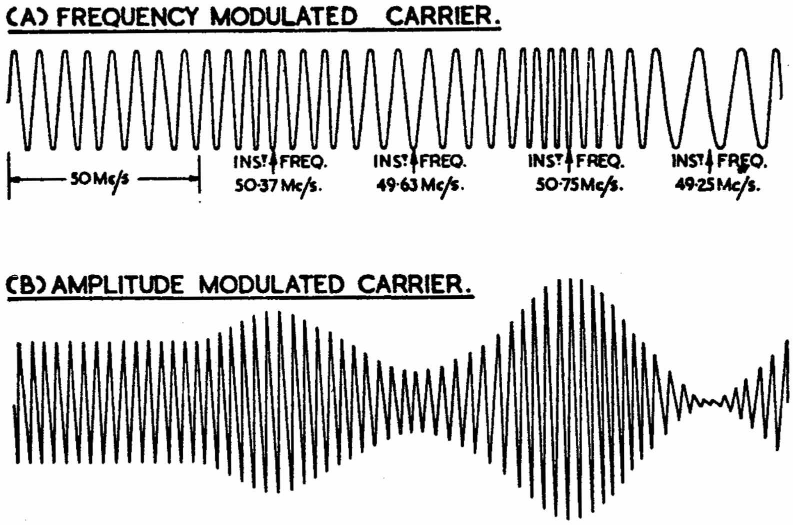
NORTHERN ALBERTA INSTITUTE OF TECHNOLOGY
EDMONTON ALBERTA

ELECTRICAL DEPARTMENT

THE FREQUENCY MODULATION OF A CARRIER WAVE

Before commencing a detailed examination of the structure of a frequency modulated wave, it will be found helpful to have a general idea of the way in which intelligence may be conveyed by a carrier wave-form. The two principal methods by which a wave may have a second signal impressed upon it are indicated in Fig. 1. The first diagram illustrates the application of frequency modulation to the carrier, whilst the second depicts the effect of amplitude modulation. In both cases the same modulating audio signal is applied - two cycles of a sine wave-shape. The amplitude of the first cycle is such that it results in 50 per cent modulation, and that of the second cycle in the maximum permissible modulation; that is to say 100 percent. In order to recover this audio signal wave at the receiver it is necessary, in the case of frequency modulation, to provide a demodulation circuit (or discriminator), in which the audio output voltage is directly proportional to the frequency variations of the carrier. In the case of amplitude modulation the detector output voltage must be proportional to the changes in carrier amplitude.


(A) FREQUENCY MODULATED CARRIER.

(B) AMPLITUDE MODULATED CARRIER.

Fig. 1 - The general nature of a frequency modulated carrier is compared with that of an amplitude modulated carrier.

With the aid of Fig. 1 it is also possible to make a number of deductions relating to the general nature of a frequency modulated carrier. In the first place, the carrier is steady at its mean or unmodulated frequency until modulation commences. It then swings above and below its mean frequency. The number of excursions which it makes on either side of this mean frequency is directly governed by the frequency of the modulating signal.

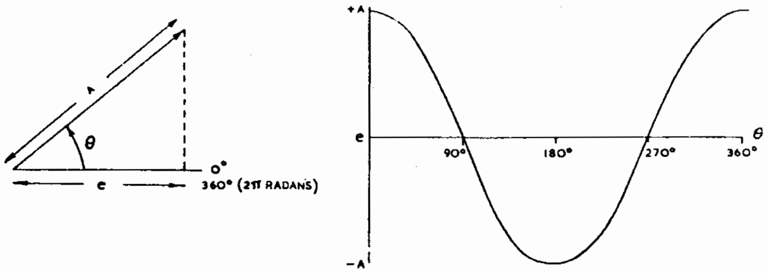


Fig. 2 - A simple alternating wave may be represented by the equation
$e = A \cos \theta$

The extent of the frequency swing is directly proportional to the amplitude of the modulating signal. It should be particularly noted that the actual frequency shift has no connection with the frequency of the modulating wave, but is entirely dependent upon i s amplitude. One of the most important points which should be brought out at this stage is the fact that the carrier amplitude remains constant regardles of the modulation depth.

In summing up, the general nature of a frequency modulated transmission may be defined as one in which there is no amplitude modulation of the carrier, and in which its frequency faithfully follows the amplitude changes of the modulating wave-shape. In the case of amplitude modulation the carrier amplitude is varied without producing any frequency variation. The amplitude changes are in direct proportion to the modulating signal's amplitude and frequency.

MODULATION

Having now outlined the general form of amplitude and frequency modulated carriers, it is possible to pass on to a more detailed consideration of the whole process of modulation.

The modulation of a wave may be defined as the process by which some characteristic is altered in accordance with the variations of a second signal, such as the voltage fluctuations associated with speech, music, television or telegraph signals. It is proposed, firstly, to establish which of the basic characteristics of a wave can be modulated.

A simple alternating voltage may be represented by the equation:

$$e = A \cos \theta, \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; (1)$$

where e = the instantaneous voltage amplitude of the wave;
A = the peak voltage amplitude of the wave;
θ = the instantaneous value of the angle of rotation of the wave vector. This may also be expressed as

$$\theta = \int_0^t w\,dt, \; . \; . \; . \; . \; . \; . \; . \; (2)$$

where $w = \frac{d\theta}{dt}$ is the instantaneous value of the angular velocity of rotation of the wave vector.

It may therefore be said that

$$e = A \cos \int_{o}^{t} wdt. \; . \; . \; . \; . \; . \; . \qquad (3)$$

The two basic methods of modulation can be identified from this equation as:

1. Amplitude modulation in which A is varied, and w is constant. In this case, expression (3) becomes

$$e = A(t)\cos(wt + \phi), \; . \; . \; . \; . \qquad (4)$$

where A(t) indicates that A varies with time; $\phi = \theta$ at t = 0.

2. Angular modulation in which w is varied, and A is constant. In this case expression (3) becomes

$$e = A \cos \int_{o}^{t} w(t)dt, \; . \; . \; . \; . \; . \qquad (5)$$

where w(t) indicates that w varies with time.

These two basic modulation groups are in turn divided into a number of different sub-groups each with its particular merits and characteristics. In the first group there is simple amplitude modulation and all the various forms of pulse amplitude modulation. Falling within the second group are phase and frequency modulation -- both being special forms of angular modulation.

AMPLITUDE MODULATION

Let is be supposed that a regular periodic change is made about the mean carrier amplitude, at a rate which is slow compared with the carrier frequency. The signal, and it should be observed that the term signal is used in this chapter to denote the modulating wave-form and not the complete modulated carrier, can be expressed as:

$$A_a \cos w_a t,$$

where A_a = the peak signal voltage;

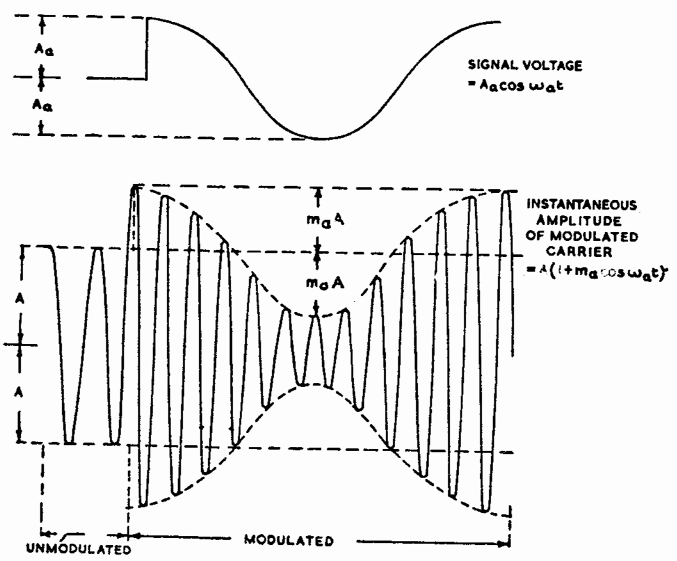
$w_a = 2\pi f_a$, where f_a is the modulating signal frequency;

$w_a t$ = the signal voltage vector rotation measured in radians.

If now the percentage amplitude modulation is made equal to a modulation factor m_a , multiplied by 100, then it follows from the definition of an amplitude modulated wave that $m_a A \; A_a$.

Under these conditions A(t) in equation (4) becomes

$$A(1 + m_a \cos w_a t) \; . \; . \; . \; . \; . \; . \; . \; . \; . \qquad (6)$$



Aa
Aa
SIGNAL VOLTAGE = Aacos wat
maA
maA
A
A
INSTANTANEOUS AMPLITUDE OF MODULATED CARRIER = A(1+macoswat)
UNMODULATED
MODULATED

Fig. 3 - Carrier amplitude modulated with a cosine wave signal. A modulation factor $m_a = 0.5$, results in 50 per cent modulation.

It will be seen that this indicates a periodic amplitude change about the value of the unmodulated carrier amplitude A, the extent of this change being determined by the modulation factor m_a. If A had been merely modified by ma cos $w_a t$, this would have indicated a change about a zero datum line.

By combining expressions (4) and (6), and taking $\phi = 0^\circ$, an expression for an amplitude modulated carrier is obtained.

$$e = A \cos wt(1 + m_a \cos w_a t) \; . \; . \; . \; . \; . \qquad (7)$$

This formula indicates that the wave consists of a high-frequency carrier, A cos wt, which is constant in frequency, but which is varied in amplitude in accordance with the signal wave, about the mean carrier amplitude A.

Expression (7) can be expanded to give the full spectrum distribution as follows:

$$e = A \cos wt + Am_a \cos w_a t \cos wt$$

$$= A \cos w_a t + \frac{m_a A}{2} \cos(w - w_a)t + \frac{m_a A}{2} \cos(w + w_a)t \qquad (8)$$

From this it will be seen that the same modulated carrier may also be considered as being built up of a spectrum of constant amplitude, constant frequency waves. This spectrum consists of the original carrier, A cos wt, and two sets of high-frequency waves,

$\frac{m_a A}{2} \cos(w - w_a)t$ and $\frac{m_a A}{2} \cos(w + w_a)t$. known as the side bands, and spaced f_a cycles on either side of the carrier. The amplitude of these side bands will be dependent on the modulation factor m_a, and will at 100 per cent modulation (i.e. when $m_a = 1$) reach a maximum of one-half the carrier amplitude.

The magnitude of the modulated wave at any instant is given by the sum of the projections on the reference axis $\theta = 0$ of the three vectors corresponding to the components of the wave, as shown in Fig. 4(a). The instantaneous wave magnitude can also be found by considering the projections of the side band vectors on the carrier vector. This leads to the vector diagram of Fig. 4(b); the upper side band vector rotates in the positive (anti-clockwise) direction relative to the carrier vector, whilst the lower side band rotates in the negative direction. The instantaneous magnitude of the carrier vector is thus $A(1 + m_a \cos w_a t)$ as given in expression (8).

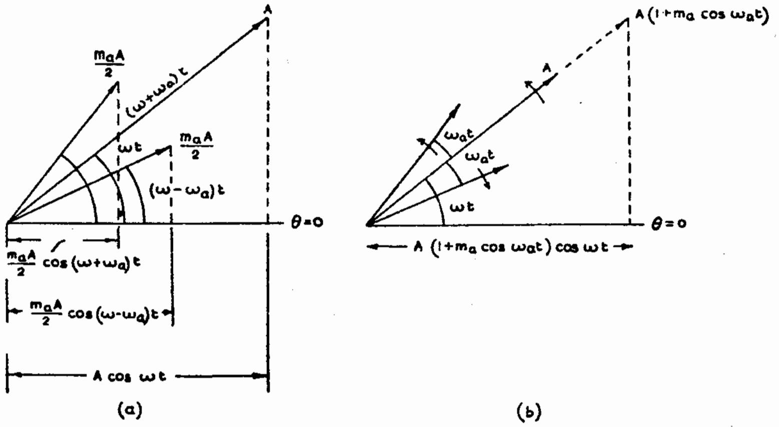


Fig. 4 - Diagram (a) shows the wave magnitude as the sum of the projection of the side band and carrier vectors on the axis $\theta = 0$. Diagram (b) shows the variation of the carrier vector magnitude as the sum of its unmodulated magnitude and the projections of the side band vectors upon it.

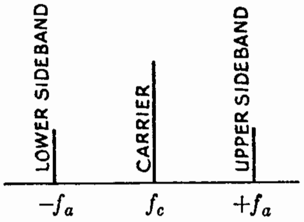


Fig. 5 - The side band spectrum of simple amplitude modulated wave.

The total radiated power contained in the side bands at 100 per cent modulation will be half the carrier power, which remains unchanged under all conditions. It will be shown in the following section that matters are entirely different for all forms of angular modulation including, of course, frequency modulation, where the total radiated power remains constant, and a large proportion of this power is contained in the side bands. It is even possible for the carrier amplitude to fall to zero. It is this important difference which makes a frequency modulation transmitter so much more efficient than its amplitude modulation counterpart.

ANGULAR MODULATION

The general expression for all forms of angular modulation is given by (5),

$$e = A \cos \int_0^t w(t)dt,$$

where $w(t)$ is the instantaneous value of the angular velocity of the wave vector. This can be expressed as the sum of two components, one constant and equal to the angular velocity (w_c) of the unmodulated carrier vector, and the other varying with time, related to the modulating signal amplitude. Then

$$w(t) = w_c + w_1(t) . \; . \; . \; . \; . \; . \; . \; . \; . \; . \quad (9)$$

The actual value of $w_1(t)$ will be considered in detail in the discussion of the various types of angular modulation.

Combining expressions (5) and (9), the instantaneous value of the wave amplitude is given by

$$e = A \cos \int_0^t \left\{ w_c + w_1(t) \right\} dt$$

$$= A \cos \left\{ w_c t + \int_0^t w_1(t) dt \right\} . \; . \; . \; . \; . \quad (10)$$

$$= A \cos \left\{ w_c t + \phi(t) \right\} \; . \; . \; . \; . \; . \; . \quad (11)$$

where $$\phi(t) = \int_0^t w_1(t) dt , \; . \; . \; . \; . \; . \; . \; . \; . \; . \quad (12)$$

is the instantaneous value of the wave phase angle ϕ, the angle between the modulated carrier vector and the mean or unmodulated carrier vector.

From expression (11), it can be seen that angular modulation can also be defined in terms of variation of the wave phase angle ϕ. If the wave frequency is made to vary directly with the amplitude of the modulating signal, frequency modulation results; if the wave phase angle is made to vary directly with the amplitude of the modulating signal, phase modulation results. Before discussing the forms of angular modulation in particular, it is necessary to elaborate on the meaning of wave frequency and phase angle, and the relationship between the two.

WAVE FREQUENCY AND PHASE ANGLE

The frequency of a wave is normally defined as the number of rotations of the wave vector (cycles) in a given period of time, generally expressed in cycles per second, or multiples of this unit. Where, however, the wave angular velocity is not constant, as in the case of angular modulation, the frequency as estimated by the number of vector rotations in a period of time yields only an average value. In order to define the instantaneous value of the wave frequency, the angle swept out per rotation (2π radians) must be divided by the instantaneous value of the wave vector velocity. This then, is the time of rotation the vector wave would have if the instantaneous value of the angular velocity $w(t)$ were maintained over a period; consequently the corresponding instantaneous value of the wave frequency is the inverse of this. Designating the instantaneous wave frequency $f(t)$,

$$f(t) = \frac{w(t)}{2\pi} \quad . . . . \; (13)$$

When the wave vector angular velocity has a fixed and a variable component, as defined in expression (9),

$$f(t) = \frac{w_c}{2\pi} + \frac{w_1(t)}{2\pi}$$

$$= f_c + f_1(t), \quad . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \quad (14)$$

where f_c is the carrier frequency,

$f_1(t)$ is the instantaneous frequency corresponding to $w_1(t)$,

i.e. $2\pi f_1(t) = w_1(t)$

Expression (14) states that the instantaneous value of the wave frequency shift, i.e. the departure of the wave frequency from its unmodulated value, is equal to $f_1(t)$. If, then, $f_1(t)$ is directly proportional to the modulating signal magnitude, the wave frequency shift is proportional to the modulating signal magnitude, and hence this type of angular modulation is termed frequency modulation.

The instantaneous value of the wave phase shift is defined as the angle between the instantaneous position of the wave vector and the position it would occupy if unmodulated. If this phase shift $\emptyset(t)$ as defined in expression (11) is made directly proportional to the magnitude of the modulating signal, the form of angular modulation termed phase modulation results.

The relationship between the instantaneous value of the wave frequency shift $f_1(t)$ $(f(t) - f_c)$ and the instantaneous value of the phase shift $\emptyset(t)$ can be found by combining expressions (12) and (14).

$$\emptyset(t) = \int_0^t 2\pi f_1(t)dt, \quad . \ . \ . \ . \ . \ . \ . \ . \ . \ . \quad (15)$$

or, alternatively, by differentiating (15),

$$\frac{d}{dt}\left\{\emptyset(t)\right\} = 2\pi f_1(t). \quad . \ . \ . \ . \ . \ . \ . \quad (16)$$

These expressions are of fundamental importance, since they show that frequency shift and phase shift are inseparable, and the relationship between them. Expressed in words, it may be stated that the instantaneous value of the wave frequency shift is equal to $1/2\pi$ times the instantaneous rate of change of phase angle.

FREQUENCY MODULATION

As stated above, if the wave frequency shift is made proportional to the modulating signal magnitude, frequency modulation ensues. With a cosinusoidal modulating signal, the resultant wave will have alternate "compressions" and "rarefactions", to borrow from the sound-wave analogy. The degree of "compression" and "rarefaction" will be proportional to the amplitude of the modulating signal whilst the occurrence of the "compressions" and "rarefactions" will correspond to the signal frequency.

It is convenient at this point to define the terms used in connection with frequency modulation; in particular the meaning assigned to frequency shift, frequency swing and frequency deviation. The term frequency shift is used to describe the departure of the signal frequency from its unmodulated value. The term frequency swing is reserved for the maximum value of frequency shift with a sinusoidal input signal, i.e. the frequency swing corresponds to the amplitude of the modulating signal. The term frequency deviation is a parameter of a given transmitting system, and is the maximum value of frequency shift permitted; this point is discussed further later.

If the signal applied to the input of the modulating system is $A_a \cos w_a t$, and b is a constant, equal to the frequency shift occurring per volt of applied signal,

$$f_1(t) = bA_a \cos w_a t. \quad . \; . \; . \; . \; . \; . \; . \; . \; . \qquad (17)$$

From expression (14) $f_1(t) = f(t) - f_c = w_1(t)/2\pi$ and combining this with expression for a frequency modulating wave becomes

$$e = A \cos\left\{ w_c t + \int_o^t 2\pi b A_a \cos w_a t dt \right\}$$

$$= A \cos\left\{ w_c t + \frac{2\pi}{w_a} b A_a \sin w_a t \right\}$$

$$= A \cos\left\{ w_c t + \frac{bA_a}{f_a} \sin w_a t \right\} \; . \; . \qquad (18)$$

since $2\pi f_a = w_a$.

This expression may be rearranged into a more general form by eliminating b and A_a. These terms are associated with the modulating sustem, and it is more convenient generally if the wave frequency swing is introduced. If f_s is the frequency swing corresponding to the amplitude of the modulating signal, $f_s = bA_a$, expression (18) can be written as

$$e = A \cos\left\{ w_c t + \frac{f_s}{f_a} \sin w_a t \right\}. \; . \; . \; . \qquad (19)$$

By analogy with the case of amplitude modulation, it might be expected that 100 per cent modulation would occur when the maximum value of the frequency swing equalled the unmodulated carried frequency; in this case, the carrier frequency would be swept between the limits 0 and $2f_c$. Such a system is, however completely impracticable.

In practice, an arbitrary upper limit f_d is set for the frequency swing and this is called the frequency deviation. This upper limit may be considered the equivalent of 100 per cent modulation. The choice of this limit is governed by two primary factors, signal to noise ratio and the band-width required for transmission. As will be shown later, the limit is required to be as high as possible to secure a good signal to noise ratio. The limit is required to be as low as possible to reduce the band-width required for transmission. The compromise value ganerally adopted for broadcasting systems is 75 kc/s; for communications systems this is often reduced to 15 kc/s.

Since f_d corresponds to $A_{a\ max}$, the maximum amplitude of the modulating signal, it is possible to introduce a modulation factor defined by

$$m = \frac{A_a}{A_{a\ max}} = \frac{f_s}{f_d} \quad . . . . . . \qquad (20)$$

and combining this expression with expression (19),

$$e = A \cos \left\{ w_c t + \frac{mf_d}{f_a} \sin w_a t \right\} . . . \qquad (21)$$

PHASE MODULATION

If, as stated above, the wave phase angle is made directly proportional to the modulating signal amplitude, phase modulation ensues. If a cosinusoidal modulating signal is considered, the wave vector will swing about its mean or unmodulated position in such a manner that the instantaneous value of the angle between the vector and its unmodulated position is proportional to the modulating signal magnitude. The frequency of the fluctuations about the mean position will be equal to the frequency of the modulating signal. With a constant frequency input, the angular deviations increase linearly with the modulating signal amplitude. If the signal applied to the input of the modulating system is $A_a \cos w_a t$, and b_1 is a constant, equal to the phase shift in radians per volt of applied signal,

$$\phi(t) = b_1 A_a \cos w_a t. . . . . \qquad (22)$$

Combining this expression with expression (11), the expression for a phase modulated wave becomes

$$e = A \cos \left\{ w_c t + b_1 A_a \cos w_a t \right\}. . . . .(23)$$

By analogy with the frequency modulation case, $b_1 A_a$ may be replaced by $m\phi_d$, where ϕ_d is the phase shift produced by the maximum amplitude of the modulating signal, and m is the modulation factor defined by $m = A_a/A_{a\ max}$. Whence

$$e = A \cos \left\{ w_c t + m\phi_d \cos w_a t \right\} ..... \qquad (24)$$

RELATIONSHIP BETWEEN FREQUENCY AND PHASE MODULATION

It was shown in expressions (15) and (16) that any frequency shift of a wave is accompanied by phase shift, and conversely. Thus a frequency modulated wave may be considered in terms of the phase shift of the carrier vector; similarly, a phase modulated wave may be considered in terms of the wave frequency shift.

Consider firstly a frequency modulated wave. The instantaneous value of the phase shift can be seen directly from expression (21) to be

$$\phi(t) = \frac{mf_d}{f_a} \sin w_a t \quad . . . . . . \qquad (25)$$

This expression shows that, with a constant amplitude modulating signal, i.e. m constant, the wave phase shift is swept between the limits inversely proportional to f_a in contrast to the analagous case in phase modulation, where the limits are constant. It also shows that the instantaneous value of the phase shift for a frequency modulated wave is in quadrature with the modulating signal magnitude. Both of these effects are due to the fact that the phase shift is proportional to the integral of the frequency deviation. If the signal applied to the modulating system had been made proportional to the differential coefficient of the modulating signal, the processes of differentiation and integration would nullify each other, and a phase modulated signal would result. Since the process of differentiating a signal wave-form can be achieved in practice, a frequency modulation system can be made to produce a phase modulated wave.

Considering now a phase modulated wave in terms of the accompanying frequency shift, expression (16) shows that

$$f_1(t) = \frac{1}{2\pi} \frac{d}{dt} (m\phi_d \cos w_a t)$$

$$= -\frac{1}{2\pi} m\phi_d w_a \sin w_a t$$

$$= -m\phi_d f_a \sin w_a t, \quad . \ . \ . \ . \ . \qquad (26)$$

since $2\pi f_a = w_a$.

This expression shows that, with a constant amplitude modulating signal, i.e. m constant, the frequency shift is swept between limits directly proportional to f_a , in contrast to the analagous case in frequency modulation, where the limits are constant. The expression also shows that the instantaneous value of the frequency swing is in quadrature with that of the modulating signal magnitude. These effects arise from the fact that the frequency deviation is proportional to the differential coefficient of the phase shift.

If the signal applied to the modulating system had been made proportional to the integral of the modulating signal, the processes of differentiation and integration would nullify each other, and a frequency modulated signal would have resulted. Since the process of integration of a signal can be achieved in

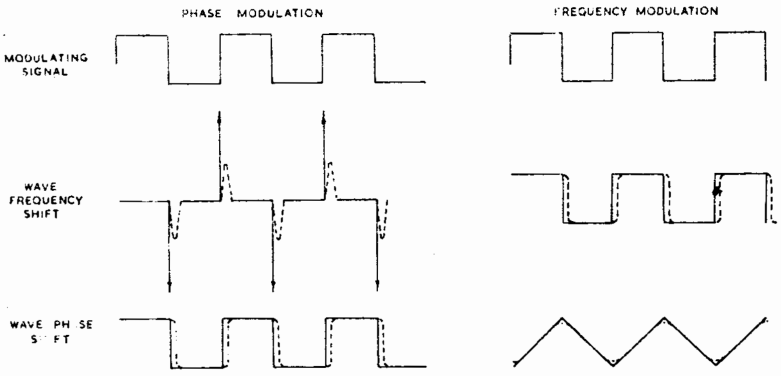


Fig. 6 - The effect on the carrier wave of a square wave modulating signal. The dotted lines indicate the practical effects obtained.

practice, a phase modulation system can be made to produce a frequency modulated wave. This fact is often utilised in practical systems.

It can thus be seen that frequency and phase modulation are very closely related; in fact, without some information as to the nature of the modulation, it is impossible to distinguish a frequency modulated wave from a phase modulated wave by inspection of the wave-form.

The differences between frequency and phase modulation can be shown clearly by considering a non-sinusoidal modulating signal. When modulation of sinusoidal type is considered, the differences are not clearly marked since the integral and differential coefficients have the same wave-shape. The differences are made most apparent perhaps by considering a rectangular wave modulation waveform, as suggested by Professor G.W.O. Howe. The resultant frequency shift and phase shift characteristics for frequency and phase modulation are shown in Fig. 6. Here the integral of the modulating signal has a triangular wave-shape, and consequently, from expression (15), the phase shift characteristic for a frequency modulated wave has this shape. The differential coefficient of the modulating signal is a series of alternate positive-going and negative-going spikes, of infinite amplitude, since the modulating signal amplitude is assumed to change by a finite amount in an infinitely short time. From expression (16), the frequency shift characteristic of a phase modulated wave also has this wave-shape.

In practice, these wave-shapes with discontinuities would be impossible to realise since they would require infinitely large band-widths for their transmission; the practical results of applying such a rectangular wave modulating signal to practical systems are indicated by the dotted lines of Fig. 6.

OTHER FORMS OF ANGULAR MODULATION

Phase and frequency modulation are not the only possible types of angular modulation; they are only two members of an infinitely large group. Another member of the group is angular acceleration modulation. Whereas in phase modulation, the phase shift is directly proportional to the modulating signal magnitude, in angular acceleration modulation, the second differential coefficient of the phase shift is proportional to the modulating signal magnitude. With an input signal A_a cos $w_a t$ applied to the modulating system, the instantaneous wave magnitude would be given by

$$e = A \cos \left(w_c t + \frac{b_2 A_a}{w_a^2} \cos w_a t, \right) \ldots\ldots \quad (27)$$

where b_2 is a constant associated with the modulating system. In this type of modulation, the phase shift is inversely proportional to the square of the modulating signal frequency. It will be noted in passing that by analogy with the name of angular acceleration modulation, frequency modulation could be termed angular velocity modulation.

It will be seen that further forms of angular modulation can be derived by making higher differential coefficients of the phase shift proportional to the modulating signal magnitude. Similarly, yet further forms could be derived by making successive integrals of the phase shift proportional to the modulating signal magnitude. However, there is no real need to consider such systems, since in practice frequency modulation is generally considered the most satisfactory type of angular modulation. This can be shown by comparison with phase and angular acceleration modulation; the successive forms suggested above merely have the relative defects of these latter types in more accentuated form.

THE RELATIVE MERITS OF FREQUENCY AND PHASE MODULATION

In view of the number of different types of angular modulation, those factors which have led to the general use of frequency modulation rather than one of the other relationships, are at least worthy of note.

There are two factors which, taken together, for all practical purposes decide the issue. Firstly, whatever method or form of modulation is employed, the limits of the channel allocated to any given transmitter must be defined in terms of frequency. The method of modulation which makes the best use of the frequency band available will therefore have much in its favour. The second deciding factor again arises from limitations which are met in practice. Up to the present all the circuits available for the demodulation of angular modulated carriers have produced an audio output voltage which is directly proportional to the variations in carrier frequency.

As the consideration of the advantages and disadvantages of frequency and phase modulation will very largely centre around these two controlling factors, it is suggested that the reader should, for convenience, also think in terms of frequency; and when considering phase or any other angular modulation visualise it as a special form of frequency modulation.

In order to assist in the building of such a mental picture, it is suggested that reference is made to the three diagrams given in Fig. 7. In these diagrams the frequency deviation resulting from 100 per cent modulation has been indicated for the three principal forms of angular modulation. It does not require a very close examination of these diagrams to show that the relationship which results in the greatest overall efficiency in the use of the frequency band employed is undoubetdly frequency modulation. By efficient use of a band, it is meant that the frequency space necessary is all employed to an equal extent in conveying the signal.

It has already been stated that the practical demodulation circuits available have a direct frequency to output voltage relationship. As most normal programme material produces maximum modulation depths over the band from 100 to 1,000 c/s, it is obvious that the demodulator circuit should be supplied with a signal which will allow it to produce its full voltage output over this region. Normally, the signal voltages over the remainder of the audio band will be of smaller amplitude. As the discriminator (the frequency modulation detector circuit) output voltage is the direct resultant of the carrier frequency deviations, it is apparent that if the full output is to be usefully employed, the modulation system adopted must be one in which this band of audio frequencies produces the maximum frequency deviation which can be permitted. Reference to Fig. 7 shows that frequency modulation alone fulfils these conditions.