Negative feedback and non-linearity

Exploring the fallacy that n.f.b. reduces all harmonics equally

by 'Cathode Ray'

MR M. G. SALEM has recently called attention (in the July issue, pp. 59-60) to the fallacy, apparently not yet extinct, of supposing that negative feedback reduces all distortion harmonics equally, by the same factor as it reduces the gain of the amplifier to which it is applied. In doing so he mentioned, in effect, that I had put an explosive charge under this particular fallacy quite a long time ago – actually April 1961 under the above title, repeated as Chapter 19 in *Essays in Electronics*. Having myself forgotten doing so, I feel confident that few other than Mr Salem are so familiar with the said work that the following revised version of it will be greeted with widespread cries of protest against excessive repetition.

Undoubtedly the first thing to learn about negative feedback is that it is never so simple as it looks. Superficial study gives one the impression that it reduces undesirable things such as distortion of all kinds and noise, mains hum, etc., dividing them by $(1-AB)$ or $(1 + AB)$, depending on the conventions adopted. Also that the input and output impedances of the amplifier to which it is applied are either decreased or increased – in the same ratio? The truth is that, even if such complications as phase shifts are excluded, none of these things is necessarily so. The effects on impedance will *not* be in the same ratio. In general, noise reduction won't be, either. Some kinds can even be increased by negative feedback. In simple cases the reducing effect on distortion is more dependable, but even there one can easily go wrong, as in the example Mr Salem pointed out. That example concerned non-linearity distortion, the effect of which is to introduce signal frequencies (harmonics and intermodulation) not present in the original. Reducing non-linearity is usually the main object of negative feedback, because that causes the most objectionable kind of distortion. No amplifier with any claim to be suitable for high-quality sound reproduction would be without negative feedback.

So let us start with a reminder of how it is commonly said to reduce non-linearity distortion. Fig. 1(a) shows an amplifier with an A-fold voltage gain. For every millivolt (say) applied to the input it gives A millivolts out. To simplify things later, we assume that the amplifier is a phase-reversing one, as indicated by the gain being shown as $-A$. Now feed back a fraction B of this output, as at (b). The voltage fed back is thus $-AB$. From the point of view of the input terminals of the amplifier the $-AB$ fed back is in opposition to the signal required between those terminals ($=1$). The signal needed between XX to maintain the amplifier signal level as before is therefore $1 + AB$, of which the $+AB$ offsets the $-AB$ fed back, leaving a net input of 1*.

Fig. 1 thus shows that negative feedback reduces the overall or gross gain of the amplifier from A to $A/(1 + AB)$ – often denoted by A'. At this point all the books mention that if the design makes AB so much larger than 1 that 1 can by comparison be neglected, A' becomes (as near as makes no matter) $1/B$. The great significance of this is that B usually depends solely on something like a potential divider that is perfectly linear, so the non-linearities involved in A are more or less removed. These and other advantages are paid for by the extra amplification needed to make AB very much larger than 1 and at the same time to ensure enough net input.

We now switch attention to the distortion created inside the amplifier by its non-linearity. It can be considered as if due to an additional input, say d millivolts; or, hopefully, microvolts. At first we might suppose that because applying negative feedback reduces the gain from A to $1/(1 + AB)$ then the legitimate signal and the distortion would both be reduced in the same ratio, so the percentage distortion would be unchanged. However, comparing (a) and (b) in Fig. 1 we see that the signal level inside the amplifier, and therefore the amount of distortion, is the same in both cases, whereas the gross signal input is much greater in (b). Therefore distortion as a percentage of the signal has been reduced by feedback in the same ratio as the gain.

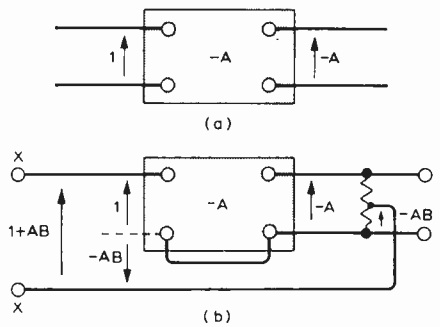

Fig. 1. *(a) represents an amplifier without feedback, and (b) the same amplifier with feedback, a fraction B of the output voltage being tapped off and returned to the input. In this case the amplifier is a phase-reversing one, so its voltage gain is shown as $-A$. The voltage fed back is therefore in opposition to the input voltage, which has to be increased accordingly; i.e., the feedback is negative.*

That is the point at which writer or reader (or both) tend to suppose that this important feedback law has been duly established and they can go on to something else. As an optional extra it may have been noted that if the gain of the amplifier is assumed to be (near enough) the same at all audio frequencies – as of course it ought to be – then the distortion harmonics and intermodulation products are all equally reduced by negative feedback.

But before hurrying on let us consider precisely what we have been meaning by A. We defined it – or, to be fair to you, *I* defined it – as the number of signal millivolts received at the output (Fig. 1(a)) for every millivolt applied at the input. But I didn't insist on millivolts, or on any particular signal level. The same A was assumed to hold good for the (presumably) much lower level of the distortion. In other words, A was assumed to be linear. That being so, it wasn't very clever to use it in a calcula-

*If no assumption is made about the polarity of the amplifier output being negative with regard to the input, the gain being called just A, then the gross input works out as $1 - AB$. This is correct for positive feedback, but for negative feedback either the amplifier or the feedback arrangement has to be phase-reversing, represented by making the value of either A or B negative, thus cancelling out the minus and giving $1 + AB$ as in Fig. 1(b). As we are considering only negative feedback, it seems rather pedantic and unnecessarily confusing to have to remember to use a double negative every time. In practice there are only (usually) two frequencies at which AB is simple plus or minus; for all the rest one has to consider other phase angles than 0 and 180 , using 'complex' algebra. But it is a *very* simple recap we are having, with a view to making just one point, not an exhaustive treatise on negative feedback.

tion concerning amplifier non-linearity. True, we guarded against complete absurdity by making the signal voltage in the amplifier the same in both Fig. 1 diagrams. But if the non-linearity is considerable, so that the distortion is a significant part of the total output, that safeguard isn't good enough. For, when the feedback is applied and reduces the distortion, the total output will be different.

The correct procedure, now that an element of doubt has been found to exist in the basis of the argument, would be to embark on a comprehensive and rigorous mathematical analysis that would cover every case. But you know me too well to expect that. Anyway, the higher the level of maths the greater the risk of going wrong or of the truth being obscured. (Mathematicians, don't bother to write to me on this, for I shall decline to answer.)

The 'line' in 'linearity' is the graph of output against input. These come in two kinds. One of them could be plotted by connecting a calibrated signal generator to the input of the amplifier and varying the signal strength there while measuring the corresponding peak or r.m.s. voltages at the output. It might look something like Fig. 2. There would

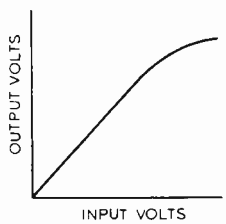

Fig. 2. *This is one kind of output/input graph, in which the voltages are peak or r.m.s. values.*

be no point in reversing the connections with the idea of extending the curve into the negative region, for its shape would necessarily be the same in reverse. The other kind, which is the one we are going to study, is seen by substituting the Y plates of a cathode ray oscilloscope for the output voltmeter, and connecting the X plates (with suitable distortionless amplification) across the input. The positive and negative half-cycles obviously swing the curve in both directions from the origin as their *instantaneous* values are shown on the screen, and their shapes are not necessarily the same.

A perfectly linear amplifier would yield a perfectly straight 'curve', as in Fig. 3(a). In the case of a power amplifier this would merely show it was being uneconomically under-driven. In a commercial world it is necessary to work up to some distortion, even though it be limited to less than 0.1%. Most amplifiers, so long as they are not over-driven, tend to show curves of two main shapes (or combinations of both), as in Fig. 3(b) and (c). The first has a

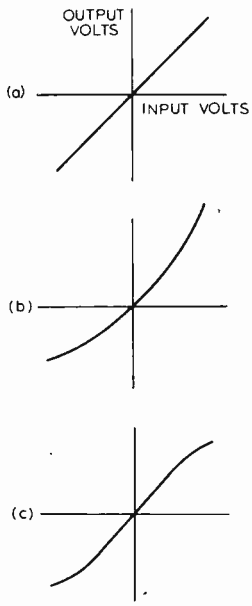

Fig. 3. *In this kind of output/input graph, instantaneous voltages are plotted. (a) is a linear (distortionless) characteristic; (b) and (c) are non-linear curves, representing respectively second and third order distortion.*

square-law term in its output/input equation, which generates a second harmonic of the signal, and second-order intermodulation. The second has a cubic term and generates third-order distortion, which sounds worse.

Now A (being output/input) is represented in these Fig. 3 diagrams by the slope of the curve. In (a) the slope is the same throughout, so A is constant and (assuming, as we usually can, that B is likewise) there need be no question as to exactly what $1 + AB$ means. But in such a situation there is no need for feedback! In (b) and (c), A is varying all the time, so one doesn't know what figure to insert for it when using the formula. We can say that Fig. 3(b) indicates a smaller A at the negative peaks than at the positive, so presumably the negative part of the curve is straightened out less by negative feedback than the positive part, but the effect on the distortion is difficult to assess without a large-scale mathematical operation. Let us see what we can do without that.

In order to find out whether the harmonic structure of the distortion (as distinct from its amount) is affected by feedback there should be no need to consider any particular practical amplifier. That is just as well, because it would be quite tricky to represent typical crossover distortion mathematically. A single transistor is easier, because it does have a Fig. 3-type graph that is a good approximation to an exponential curve, and (with suitable assumptions) the corresponding array of harmonics in the output can be derived as a basis for calculation. But why bother? Things will be much easier and clearer if (at least for a trial) we assume we have a hypothetical amplifier with a pure square-law characteristic, like Fig. 3(b), and plotted quantitatively as Fig. 4, using the equation.

$$v_o = 100v_i + 1000v_i^2$$

where v_o is the instantaneous output voltage and v_i the sinusoidal input voltage. This gives the amplifier a gain of 100 as regards the fundamental.

A simple calculation shows that with a peak v_i of 0.04V the $1000v^2_i$ term causes 20% second-harmonic distortion. We can do it graphically by drawing a straight line joining the tips of the curve, noting how far up the v_o axis it comes (1.6V in this case) and lowering the line half the distance. It is then the linear part of the characteristic responsible for the fundamental, shown (dotted) as a pure sine wave in Fig.5 (a) The actual amplifier curve I have plotted in Fig. 4 is 0.8V lower at zero v_i and 0.8V higher at positive and negative peaks. The points can be transferred to Fig. 5(a), and when joined up by the full line show what comes out of the amplifier when 0.04V peak is put in. The difference between this and the fundamental has been plotted below, (b), and is clearly a second harmonic. Both Fig. 4 and Fig. 5 show that its peak value is 0.8V, which in relation to the fundamental's 4V is 20%.

Anyone with the most elementary knowledge of the differential calculus will realize that the easiest way of finding the slope (which is A) at any point on the Fig. 4 curve is to differentiate its equation, thus:

$$A = \frac{dv_o}{dv_i} = 100 + 2000v_i$$

So at zero v_i it is 100, which is what one would expect, since an input confined to very small values of v_i would yield neligible distortion, and 100 is the slope of the fundamental line, corresponding to an amplification of 100. At the

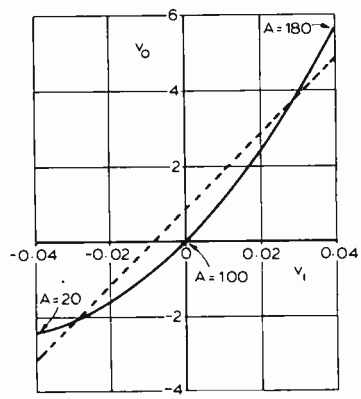

Fig. 4. *The full line is a graph of the Fig. 3 (b) type. The broken line shows its fundamental part; the vertical difference between the two represents second-harmonic distortion, as shown in Fig. 5.*

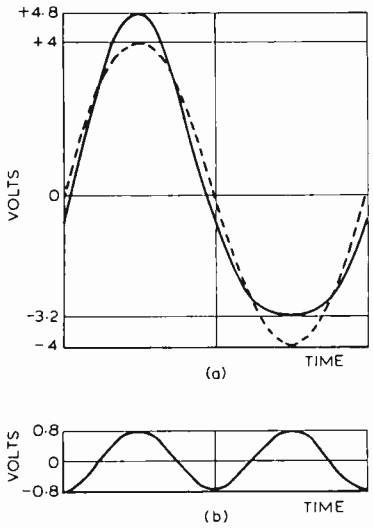

Fig. 5. *The full line in (a) shows the output waveform of an amplifier with the characteristic shown in Fig. 4, when the input is a pure sine wave. The broken line is the fundamental part, corresponding again to Fig. 4. The difference between the two, shown by itself at (b), is the second harmonic.*

positive peak it is 100 + 80 = 180 and at the negative peak it is 100 − 80 = 20. So 20% distortion, which is not as horrible as you might expect, if it is all second harmonic, is associated here with a no less than 9 to 1 variation in amplification over each cycle of signal. We can hardly be surprised, then, if we find that negative feedback doesn't work entirely according to plan.

Perhaps the best way of seeing how it does work is to plot a with-feedback curve to compare with Fig. 4, which can be done by making a table to calculate some points. Remember, the voltage fed back at any point is equal to Bv_o, and this added to v_i gives v'_i, the with-feedback input needed.

To make it easy to compare the two curves, the v' scale of the new one should be the v_i scale of the old multiplied by as many times as v'_i must be greater than v_i to maintain the same output. A convenient figure for this, which is also reasonable for feedback practice, is 10. $(1 + AB)$ being 10, AB is 9 and B is 0.09.

(1) v_i	(2) v_o	(3) $0.09v_o$	(4) v'_i
0.01	1.1	0.099	0.109
0.014	1.596	0.1436	0.1576
0.02	2.4	0.216	0.236
0.03	3.9	0 351	0.381
0.04	5.6	0.504	0.544
−0.01	−0.9	−0.081	−0.091
−0.02	−1 6	−0.144	−0.164
−0.03	−2.1	−0 189	−0.219
−0.04	−2.4	−0.216	−0.256

Column (1) contains a selection of points covering the peak-to-peak swing of v_i. Column (2) contains the corresponding output voltages calculated from the equation, which were needed for plotting Fig. 4. Column (3) shows the voltages fed back, equal to $0.09v_o$. Lastly column (4), which is got by adding (3) to (1), shows the input required at XX in Fig. 1(b) to maintain the same output (2) as before.

Plotting Fig. 6 from columns (2) and (4) we are at once impressed by the success of negative feedback in straightening out the amplifier curve. It is now hardly distinguishable from a straight line, especially on the positive side.

Becoming a little more critical, we note that we need considerably more than 10 times the former peak input; to be exact, 13.6 times. But 10 was calculated on the basis of $A = 100$, whereas we have already noted on Fig. 4 that A varies from 100 to 180 during the positive half-cycle, and if we calculate the average multiplier for this range of values of A we find that it is 13.6. Rather than find fault here we might thank feedback for raising the positive fundamental peak output from 4V with 20% distortion to 5.5V with about 1½% distortion.

On the other hand any satisfaction that might at first be derived from seeing that the input needed for the negative peak has been increased only 6.4 times is damped by the unfortunate accompanying fact that the fundamental negative peak has been reduced from 4V to about 2.5V. And of course a 5.5V positive peak is no good with a 2.5V negative peak – unless use of the amplifier is confined to rather unusual waveforms.

It seems, then, that if at least our original 4V peak sine-wave output is to be maintained it will be necessary to bring up the negative input, as we would be able to do, seeing that we were prepared to find at least ±0.4V. To see what we get we shall have to extend our plots in the negative direction. If we proceed to calculate column (2) we find that beyond $v_i = -0.05$V a complication sets in; increasing $-v_i$ reduces $-v_o$, making the curve bend up. This is because the curve is derived from the equation for A, which makes A negative if v_i is more negative than −0.05V.

In a real amplifier, however, the decline of its gain to zero during the signal cycle is normally due to the cutting off of one or more transistors. In the usual push-pull configuration – in which distortion is largely third-order – the gain recovers as the signal goes more negative; indeed, if the biasing is correct it shouldn't fall off in the first place. But we are considering a square-law amplifier, to which the nearest practical approximation is a single-ended type, which cuts off altogether if the signal goes too negative. So a realistic procedure will be to continue the curve horizontally to the left:

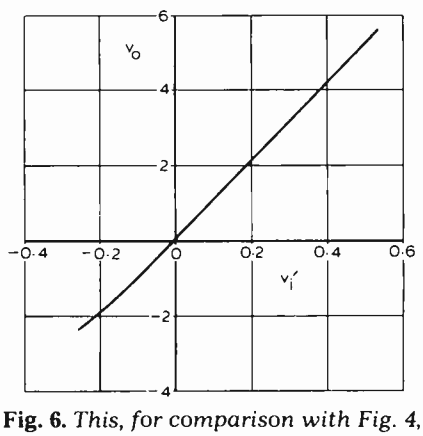

Fig. 6. *This, for comparison with Fig. 4, is the result of reducing the small-signal gain A-fold by negative feedback and correspondingly increasing the external input (v') to yield the same net input (v_i) as before.*

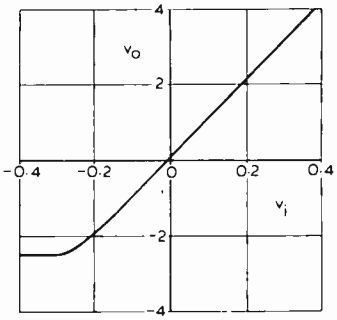

Fig. 7. *The result of further adjusting the input v_i to ±0.4V. Audibly, the result would be worse than without feedback (Fig. 4), and the output less.*

v_i	v_o	$0.09v_o$	v_i
−0.05	−2.5	−0.225	−0.275
−0.06	−2.5	−0.225	−0.285
−0.07	−2.5	−0.225	−0.295

At this rate it is obviously going to take us a long time to reach $v'_i = -0.4$, but we now have enough information to omit the intermediate stages and boldly write

$$v'_i = -0.4;\ v_o = 2.5$$

Continuing beyond our original ±0.4V (to match the ±0.04V in Fig. 4) is clearly not going to make the picture any prettier, so in Fig. 7 I have kept within those limits. Now we see the truth about negative feedback, and it doesn't look as good as we may have supposed. And if anyone is thinking I've fiddled it by arbitrarily departing from the simple quadratic equation at the negative end, I invite him to stick to the equation. The result will be even more ghastly than Fig. 7.

That is quite bad enough, for on analysing Fig. 7* I find that the fundamental output is only just over 3V peak, compared with 4V in Fig. 4 (a power reduction of 44%) and in exchange for our 20% second-harmonic distortion we have received the following mixed bag:

Harmonic	Percentage
2	13.2
3	7.4
4	3.3
5	1.24
6	0.16
7	0.83

*By the method described in M. G. Scroggie's *Radio & Electronic Laboratory Handbook*, 8th edition, Sec. 11.10.

plus undetermined amounts of higher harmonics which, judging from the sharpness of the bend in Fig. 7 and the magnitude of the 7th harmonic, are likely to be very significant aurally if not numerically. It is true that the total harmonic distortion, found by taking the square root of the sum of the squares of the above lot, is only 15.6% But if anyone thinks this is an improvement on the 20% without feedback he oughtn't to be let out alone in the hif-fi market. He would be an easy prey to the merchants, whose motive in quoting t.h.d. figures is only too clear to those who have compared actual sound reproduction with the harmonics present. Though opinions of authorities differ as to the factors by which harmonics higher than the second should be multiplied to give some idea of their relative unpleasantness, the most conservative suggest (without necessarily admitting that it is adequate) a weighting factor equal to half the harmonic order. For instance, the 0.83% 7th harmonic would have to be multiplied by 3½, raising it to 2.9%. D. E. L. Shorter of the BBC considered the square of this factor not excessive. That would raise the 7th-harmonic figure for comparison with the second to over 10%.

At this point a red herring labelled 'intermodulation' is almost certain to be seen crossing our path. But if any benefit is to be derived from the time you have so self-sacrificingly spent in following me thus far, I advise that we refrain from spending any more in chasing after it. No doubt we know that the products of intermodulation, being in general not musically related to the tones present in the original sounds, are more objectionable than at least the lower harmonics, which are; but it doesn't follow that one must insist on intermodulation data and refuse harmonics as worthless substitutes. For, when measured under comparable conditions, harmonic percentages are more or less proportional to intermodulation percentages, so can be used as comparative indexes of intermodulation, easier to measure. And anyway, in this case we are getting the higher harmonics, which are discordant in their own right.

Another possible red herring is one that isn't nearly as fresh as it is often made out to be by means of new-fangled terms such as transient intermodulation distortion and slewing-rate distortion. It is in fact many years old, and although it too is an undesirable product of ill-designed negative feedback it also is an avoidable one, not directly related to the present subject.

Getting back to our uneasy contemplation of Fig. 7, we see that there is nothing for it but to reduce the input signal until the sharp bend is cleared; say ± 0.25V peak. The output, which by then is nearly all fundamental, is barely 2.5V, or less than 40% of the power we got in Fig. 4, admittedly with lots of second harmonic too. But if we reduce the fundamental without feedback to the same level, the second harmonic comes down to 12½%, which on paper is certainly not hi-fi, but wouldn't greatly offend as many listeners as you might think.

It is now time to sum up:

(1) The common belief that negative feedback reduces non-linearity distortion in the same ratio as it reduces amplification is strictly true only if there is no non-linearity to reduce.

(2) However, provided that the original non-linearity is not so bad that the slope of the output/input curve (which is the amplification) falls seriously below the nominal value at any point within the maximum signal amplitude, the common belief is fair enough.

(3) It follows from (1) and (2) that any idea that one can sling an amplifier together any old how and pull it straight with liberal supplies of negative feedback is unsound — even apart from the practical difficulties of this treatment.

(4) While negative feedback works like a charm on amplifiers with *moderate* non-linearity, run well within their capability, it doesn't necessarily increase the amount of power that can be drawn; on the contrary, it may reduce it.

(5) In any case, once the signal amplitude runs past the nearly-undistorted limits, it abruptly becomes very distorted, not only as regards quantity but even more as regards quality. In other words, even a moderately overloaded amplifier sounds a lot worse with feedback than without.

(6) The fact that hi-fi fans insist, especially in America, on vast numbers of output watts being available, in spite of the surprisingly small *average* power needed for even quite loud reproduction, is thus explained.

(7) The fact that demonstrations of 'hi-fi', unless conducted by masters of the art, are usually such painful experiences, is also explained. The demonstrator so often doesn't reckon that he is doing his job if the output falls below the maximum rating.

Except by dividing signal voltages by 10 in order to be more appropriate for modern transistors than were those in the valve version of 1961, and writing a new introduction on Fig. 1, I have followed much the same lines as in the original and have arrived at the same conclusions. Present readers will no doubt be thinking I ought to have reduced the distortion figures by a factor of at least 10 to be more in accord with present-day amplifiers. But it must be remembered that, with the larger amounts of feedback now used, its effects on overloading can be even worse than are shown here, intentionally exaggerated though they were to get the message across. This has been dramatically confirmed as recently as the July 1978 issue, where on p.57 James Moir showed a curve which clearly illustrates my very point — that distortion without feedback is, at a certain output level, suddenly and vastly overtaken by distortion with feedback.

I hope that, by confining the no-feedback distortion to only one harmonic, I have left no room for the fallacy that all distortion harmonics are necessarily reduced by negative feedback in the same ratio as the gain – or even at all, since we have seen that many harmonics can actually be created by feedback that were not there without it. □

LITERATURE RECEIVED

Video display unit ZIP-64 from Data Dynamics is said to offer low cost with high performance. A leaflet can be had from Data Dynamics at Data House, Springfield Road, Hayes, Middlesex WW412

P.r.o.m. programming equipment made by Data I/O and a large list of p.r.o.m.s from twenty suppliers is presented in a leaflet from Microsystem Services, Duke Street, High Wycombe, Bucks. WW413

Illuminated push switches illustrated and described in 28-page catalogue from Licon, Norway Road, Hilsea Industrial Estate, Portsmouth PO3 5HT WW414

Power supplies and components for use with equipment vulnerable to transients and poor line regulation and in conditions where a supply must not be broken are all described in the Topaz catalogue from Euro Electronic Instruments Ltd, Shirley House, 27 Camden Road, London NW1 1YE WW415

Single-board computers in the Intel iSBC range of o.e.m. equipment have been summarized by Rapid Recall in a pocket guide, obtainable from Rapid Recall at 9 Betterton Street, Drury Lane, London WW416

Turntables from Collaro are updated and described in leaflets from Magnavox Electronics Company Ltd, By-pass Road, Barking, Essex 1G11 0TF WW417

Picoammeter from Keithley, Model 480, is discussed in general and specified in a brochure from Keithley Instruments Ltd, 1 Boulton Road, Reading RG2 0NL . WW418

"DC Motors, Speed Controls, Servo Systems" is the title of a 500-page handbook from Electrocraft. It is available at £3 from Unimatic Engineers Ltd, Granville Road Works, 122 Granville Road, Cricklewood, London NW2 2LN.

Audio kits from Powertran are illustrated, described and priced in a catalogue obtainable from Powertran Electronics, Portway Industrial Estate, Andover, Hants SP10 3NN. WW designs offered include the Linsley Hood cassette deck, Nelson-Jones f.m. tuner, Stuart tape recorder and Linsley Hood audio oscillator WW419