Audio power amplifier design

There is nothing so practical as a really good theory — LUDWIG BOLTZMANN

by **Peter J. Baxandall** B.Sc.(Eng.), F.I.E.E., F.I.E.R.E.

Articles describing particular amplifier designs, or advocating specific solutions to design problems, abound in the literature, and it is evident that some quite conflicting views exist on certain topics — for example, concerning the amount of negative feedback that should be used. The present approach is of a fairly broad nature, and aims to elucidate and compare various familiar and unfamiliar circuit techniques in such a way that their advantages and disadvantages may be clearly and logically appreciated.

IN EXPLOITING the very great virtues of negative feedback, the problems and difficulties that arise are largely those associated with obtaining adequate stability margins under all conditions of operation. In a.c. coupled amplifiers, there are stability problems at both low and high frequencies, but the elimination of output transformers, together with the adoption of d.c. coupled circuitry in most modern designs, has virtually removed the low-frequency problems.

Negative feedback and slew-rate limits

Other things being equal, the larger the amount of overall negative feedback applied to an amplifier, the lower will be the distortion. However, other things are quite likely not to be equal, since, to achieve stability, it is usually necessary to introduce elements which start attenuating the forward gain, with rising frequency, at a frequency which has to be made lower and lower as the amount of overall feedback is increased. If *unsuitable techniques* are used for effecting this attenuation, increased distortion will be generated in the forward path of the amplifier *at high frequencies*, to an extent which may more than offset the advantages of the increased feedback. Indeed, drastic high-frequency internal overloading may occur, and once this has happened, the overall feedback is powerless to preserve the wanted output waveform.

The rudimentary amplifier circuit shown in Fig. 1 will serve to illustrate the point. Here the capacitor C attentuates the gain with rising frequency by making Tr_2 function as a Blumlein integrator. The current, I, supplied by the first stage includes, in addition to a component flowing to Tr_2 base, a component much larger at high audio frequencies flowing to C. At such frequencies, and with Tr_2 producing a large output voltage swing, the current demanded by C may severely tax the output capability of Tr_1 stage, and may, in the limit, cause Tr_1 to overload, i.e. cut off during part of the cycle. Whether or not this will happen can be determined quite simply, on a sine-wave basis, by calculating the current in C, which is, nearly enough, V_{out}/X_c. If the peak value of this current exceeds the d.c. working current of Tr_1, gross distortion will occur. Thus the critical condition for the onset of such distortion is

$$I_{dc} = \hat{V}_{out} \times 2\pi fC \qquad (1)$$

This relationship may be rearranged to give a convenient formula for the critical sine-wave frequency, f_{crit}, above which gross distortion sets in no matter how much overall feedback there is. Thus

$$f_{crit} = \frac{I_{dc}}{2\pi C\hat{V}_{out}} \qquad (2)$$

It is customary nowadays, in the above context, to employ the slew-rate concept, though it is by no means essential to do so. This concept has long been familiar to workers in other fields, particularly those of servo-mechanisms and radar. As applied to amplifier circuits, the basic relationship is simply that, for a capacitor

$$dv/dt = i/C \qquad (3)$$

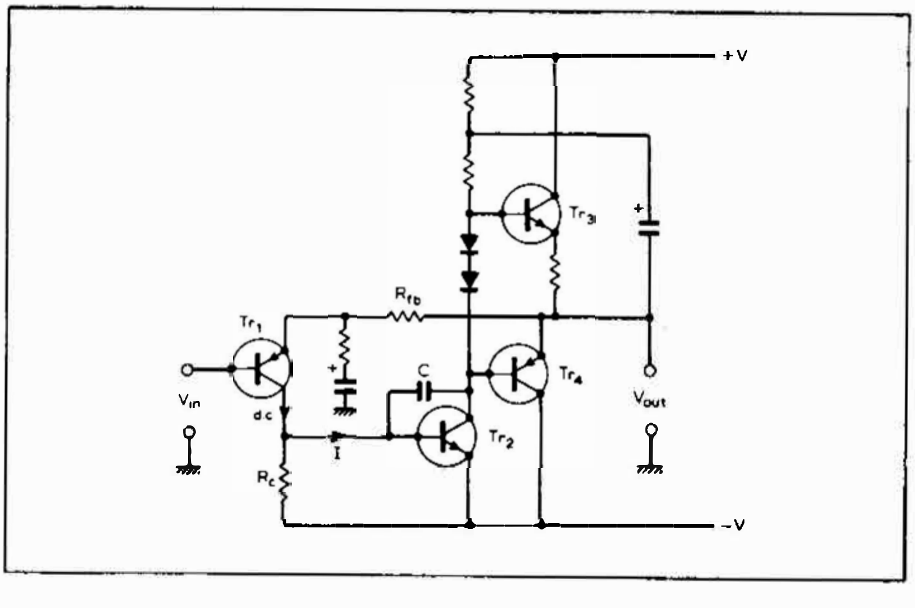
Thus, with reference to Fig. 1 again, suppose the transistor Tr_1 is briefly cut off; then a current approximately equal to I_{dc} is left flowing in R_c and most of this also flows in C, producing a positive-going rate of change of output voltage

$$[dv_{out}/dt]_{max\ poss} = I_{dc}/C \qquad (4)$$

This is called the *output slew-rate limit* of the amplifier, or sometimes, in commercial practice, just the slew-rate. With the single-ended input stage of Fig. 1, the slew-rate limit for negative-going outputs will be much more rapid than the above, because Tr_1 can turn on much more current than it can turn off. But when a balanced long-tailed-pair input stage is used, as in most integrated-circuit operational amplifiers, the slew-rate limits in the two directions will be approximately equal.

The relationship (4) applies whatever the signal waveform may be. If, at any instant, the demanded rate of change of output voltage exceeds this value, the amplifier will fail to follow it properly. Thus, if an amplifier has an insufficient slew-rate limit, then, every now and

Fig. 1 *Rudimentary amplifier circuit in which the capacitor C gives rise to slew-rate limiting.*



then, on fast transients particularly, the slew-rate limit will be exceeded by the programme waveform. When this occurs, the amplifier gain will fall drastically, and all components of the signal being handled at that moment will be chopped, or modulated, by the transient. This effect, well known to enlightened designers of feedback amplifiers for decades, has nowadays, of course, become known as transient intermodulation distortion or t.i.d. (sometimes t.i.m.), as a result of several papers by M. Otala. Another, more recent, related term, due to W. G. Jung, is slewing induced distortion, or s.i.d.[1,2,3].

It is of interest to obtain the relationship between the general slew-rate limit formula (4) and the conditions which apply with sine-wave input. Substituting in (2) the value of I_{dc}/C given by (4) yields

$$f_{cnt}=\frac{[dv_{out}/dt]_{max\,poss}}{2\pi\hat{V}_{out}}$$

$$\text{i.e.}\quad f_{crit}=\frac{\text{output slew-rate limit}}{2\pi\hat{V}_{out}}\qquad(5)$$

(This result can alternatively be obtained by differentiating the output voltage waveform, $v = \hat{V}\sin 2\pi ft$, and equating the peak instantaneous value of the differential coefficient to the slew-rate limit.)

In all the above, the slew-rate limit referred to is that of the amplifier *output* voltage, and this is the usual practice – especially in integrated circuit data sheets, where it is simply called the slew-rate. Thus, unless otherwise stated, slew-rate figures may be assumed to apply to the output of an amplifier. However, it is sometimes convenient to express them with respect to the input, which merely involves dividing by the amplifier's voltage gain. The corresponding equation to (5) for the input is

$$f_{crit}=\frac{\text{input slew-rate limit}}{2\pi\hat{V}_{out}}\qquad(6)$$

Consideration of (5) and (6) makes it evident that what is invariant is the quotient of the slew-rate limit and the peak sine-wave voltage at any selected point in the system. Hence, more generally,

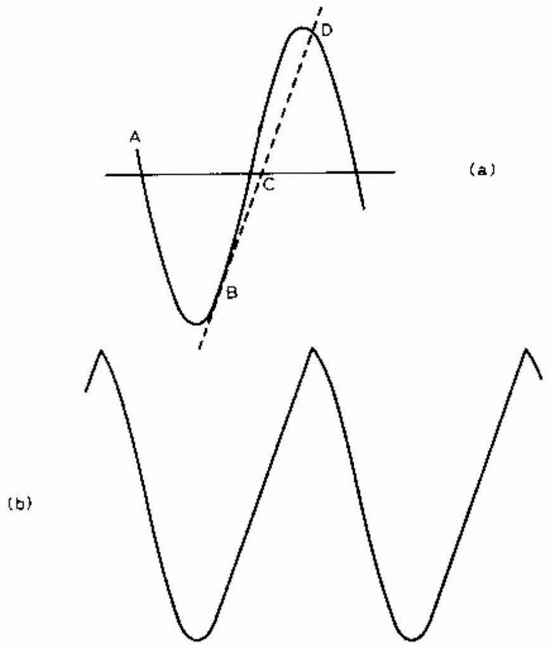
$$I=I_oe\frac{qV_{be}}{kT}$$

The peak voltage V is normally that for full output level. The quality of the slew-rate performance of an amplifier may thus be expressed by the slew-rate-limit figure given in *volts per micro-second per volt peak* of sine-wave signal. For example, f_{crit} = 20kHz corresponds to a figure of 0.126V/μs per volt peak.

It is of interest to consider what sort of output waveform would be expected from an amplifier suffering from slew-rate limitation, on sine-wave input. Suppose initially that the amplifier is basically as in Fig. 1, having a single-ended input stage which imposes a much more severe slew-rate limit for positive-going amplifier output voltage than for negative-going. Referring to Fig. 2(a), the sine-wave represents the wanted output waveform, and the broken line represents the maximum rate of change of output voltage of which the amplifier is capable, i.e. it represents the output slew-rate limit. The actual output therefore follows the wanted waveform from A to B, but after B it follows the path BCD before joining the wanted waveform again at D. The complete output waveform is thus as shown in Fig. 2(b). Fig. 3(a) shows some experimental waveforms obtained with a circuit having the basic configuration of Fig. 1, for two different degrees of slew-rate limitation overload on sine-wave input. Fig. 3(b) shows the output waveform for square-wave input, and is a typical result for an amplifier exhibiting unsymmetrical slew-rate limitation.

The waveforms of Fig. 4 were obtained using a type LM301AN integrated circuit operational amplifier as a unity-gain inverter. The 301 circuit, very broadly speaking, has a similar type of configuration to that shown in Fig. 1, but with a balanced long-tailed-pair input stage arrangement. The external stabilizing capacitor C, more often called the compensation capacitor, had a value of 30pF. It will be seen that, as expected, the slew-rate limitation is of a nearly symmetrical nature.

Fig. 2 *Diagrams illustrating unsymmetrical slew-rate limiting.*



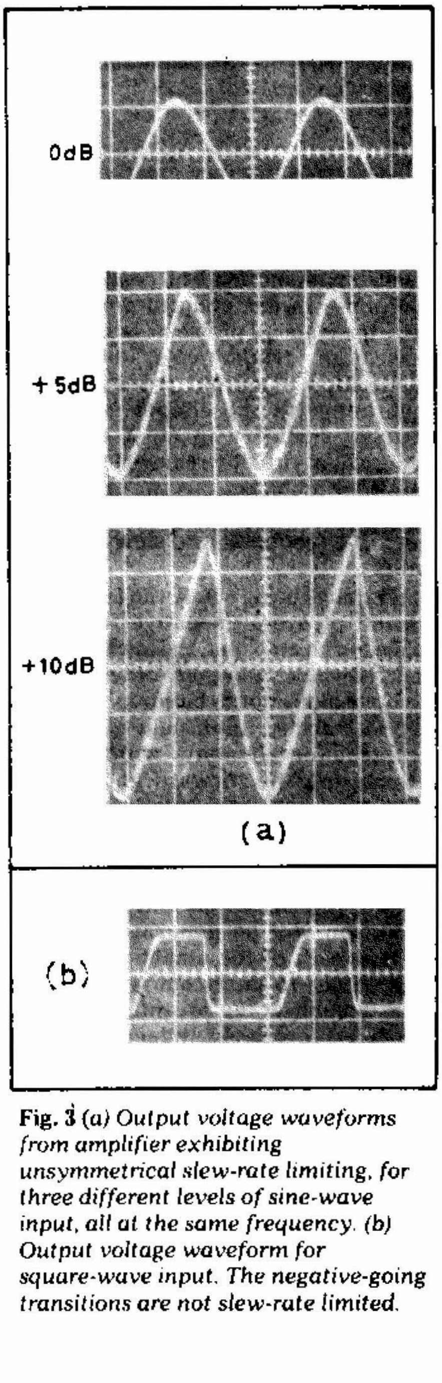
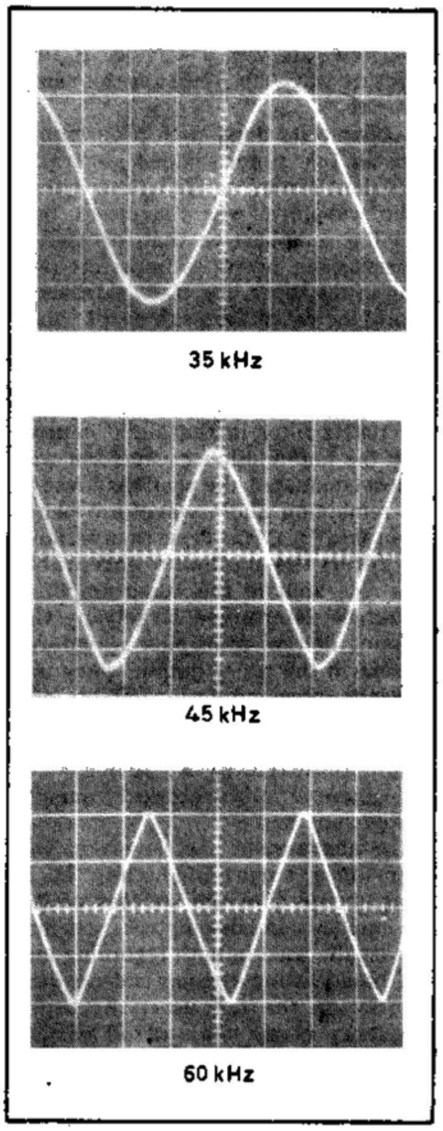


Fig. 3 *(a) Output voltage waveforms from amplifier exhibiting unsymmetrical slew-rate limiting, for three different levels of sine-wave input, all at the same frequency. (b) Output voltage waveform for square-wave input. The negative-going transitions are not slew-rate limited.*

A great deal of attention has been given to this aspect of amplifier behaviour in recent years, and while it is certainly important to avoid significant distortion of this type, the notion that it is a fairly newly-discovered form of distortion is quite unjustified. It all boils down to the fact that, to avoid unwanted intermodulation effects, a good amplifier should be able propertly to track all normal programme waveforms, whether of a sustained-tone or a transient nature, without any internal circuits overloading in the process — surely an old and familiar notion? Indeed, I cannot do better than quote Jung, who says "there is nothing new, unique, or mysterious about slew-induced or transient intermoduation distortion"[2]. It may be added, however, that since some — but certainly not all — of the earlier transistor amplifiers suffered seriously from this type of distortion, the widespread attention that has been given to it is a good thing. But removal of significant s.i.d. is not a panacea — there are also other important causes of distortion.

Fig. 4 *Output voltage from integrated-circuit operational amplifier for equal-amplitude sine-wave inputs at three different frequencies, showing slew-rate limiting. Scales: 1V/cm, 5µs/cm.*



As considered above, the slew-rate-limit mechanism sets a fairly sharply defined threshold, beyond which there is a rapid onset of gross distortion that the overall feedback is powerless to control. Below this threshold output level, which is, of course frequency-dependent, the distortion will be negligible only if there is sufficient overall feedback. Whether there is enough feedback to give this result depends on the details of the particular design, but in some instances there may not be enough. Thus it is of interest to consider the distortion mechanisms that are operative in the milder situation where drastic overloading does not occur.

Referring to Fig. 1, suppose we decide to apply 6dB more overall feedback to the amplifier by reducing R_{fb}. This is likely to necessitate doubling the value of C, for equally satisfactory stability. Thus, while we succeed in doubling the feedback loop gain at low frequencies, where C has little effect, the loop gain at higher frequencies, where C is dominant, remains as before. At a given high frequency, and a given output voltage, Tr_1 will have to supply twice the current to the doubled value of C, and the percentage second-harmonic distortion generated in Tr_1 will go up by a factor of approximately 2*. Since the amount of feedback at the high frequency involved is the same as before, the amplifier output distortion (due to distortion in Tr_1) will also be doubled.

Because of the doubling of the C value, the critical frequency for slew-rate limitation, above which full output ceases to be obtainable without drastic overload, is halved — see equation (2).

Quite frequently a long-tailed pair, or differential input stage, will be used in place of the single transistor Tr_1 shown in Fig. 1, and then, if well balanced, the dominant distortion introduced will be third-harmonic, the percentage distortion being proportional to the square of the output current [5]. (This is a characteristic of any device, e.g. a tape recorder, in which cube-law curvature is dominant.) Thus, with the low-frequency overall feedback increased by 6dB, and with C doubled as before, the third-harmonic distortion generated in the input stage will be up by a factor of 4 at high frequencies, as also will be the amplifier's output distortion due to this cause.

We thus have the situation that increasing the amount of low-frequency overall feedback, with corresponding adjustment of the stabilizing capacitor value, increases that part of the high-frequency output distortion which is due to smooth-curvature non-linearity distortion in the input stage. In many cases, below the true slew-rate-limitation overload point, this will be the main cause of distortion at high frequencies. However, with suitably modified circuit designs, to be described later, the input stage distortion may be fairly negligible.

It is interesting to consider how the above non-overloading type of distortion would be expected to vary with frequency. A long-tailed-pair input stage will first be assumed. Since, at high frequencies, the current supplied by the input stage is proportional to frequency, the percentage third-harmonic distortion generated within the stage is proportional to the square of the frequency. But because the overall-feedback loop gain is halved for each doubling of frequency, the distortion at the output of the amplifier, due to this mechanism, is proportional to the cube of the frequency. The percentage output distortion is thus proportional to $V_{out}{}^2 f^3$, as established by Jung. The corresponding result for a single-ended input stage, as in Fig. 1, is that the percentage output distortion, now mainly second-harmonic, is proportional to $V_{out} f^2$. This is because in any device in which square-law curvature is dominant, the percentage distortion is directly proportional to the output current or voltage.

It will thus be seen that a characteristic feature of distortion of the type discussed above, which occurs before the onset of true slew-rate-limitation overload, is that it increases quite rapidly with frequency. Fig. 5 shows the ideal cube-law relationship deduced above for the balanced input stage case. With a single-ended input stage, though the rise in distortion with frequency is more gradual, the magnitude of the distortion is liable to be much greater [5].

Jung calls the input-stage-originated distortion that occurs before the onset of true slew-rate limitation "Category I - slewing induced distortion", the gross distortion that occurs at higher levels being "Category II s.i.d." It is important not to let this terminology disguise the fact that Category I s.i.d. is, after all, just straightforward input-stage smooth-curvature non-linearity distortion, which may become significant at high frequencies because of the increased current demanded from the input stage and the reduced amount of overall feedback in action.

* The percentage second-harmonic distortion produced by an ideal voltage-driven transistor, having a characteristic $I = I_o \exp qV_{be}/kT$, approximately $25 \times (\hat{I}/I_{dc})$, where $\hat{I}$ is the peak value of the signal-current fluctuation and I_{dc} is the d.c. working current. Another convenient fact is that, at any working current, the percentage second-harmonic distortion is equal to the peak value, in millivolts, of the signal voltage applied between base and emitter[4,5].

Though, as shown in Fig. 5, the high-frequency distortion due to the input stage rises rapidly with the measuring frequency applied, it should not be imagined that the harmonics generated at any one measuring frequency are boosted according to their order, in any comparable manner. Consider first the effects that would occur with the overall feedback disconnected. Referring again to Fig. 1, the harmonics in the current fed by the input stage to the Tr_2 stage will be attenuated in this stage in proportion to their order, because of the integrating action of the capacitor C. Thus, with the feedback loop open, the harmonics in the amplifier output voltage, due to input stage distortion, would fall off in amplitude with increasing order at a rate 20dB/decade (6dB/octave) more rapid than that applying directly to their generation in the input stage. However, with the overall feedback loop closed, and because the amount of feedback at high frequencies falls off at 20dB/decade with increasing frequency – assuming C is the only cause of loop gain attenuation – the final output distortion spectrum will have the same relative amplitudes of fundamental and harmonics as for the input stage by itself. With a long-tailed-pair input stage, and assuming the circuit not to be operating too close to the slew-rate limit point, the dominant harmonic will be the third, the higher order harmonics decaying rapidly with increasing order. Thus the type of distortion generated is relatively innocuous compared with the worst forms of cross-over distortion. The important thing is simply to arrange the design so that the magnitude of the distortion does not become too high.

Slew-rates of programme waveforms

Gramophone records are frequently used as the programme source when subjective judgements of the performance of audio equipment are being made, so that it is of interest to know the order of slew-rate to be expected at the output of a high-grade RIAA equalized amplifier. This can easily be determined using a very simple differentiator circuit such as that shown in Fig. 6. This circuit is fed from the output of the power amplifier, and, with the values shown, gives an instantaneous output of 1 volt when the input slew-rate is 1V/μs. The objection may well be raised that the slew-rate limit may degrade the true slew-rate of the source, i.e. the pickup, but whether or not this is the case may be discovered by replacing the pickup by an oscillator and thus determining the slew-rate limit of the amplifier system. With good equipment, this will be found to be much higher than the slew-rate obtained with records.

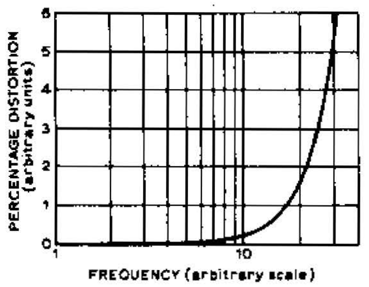


Fig. 5 *Theoretical variation of third-harmonic distortion with frequency for amplifier with long-tailed-pair input stage, when operating below the slew-rate limit.*

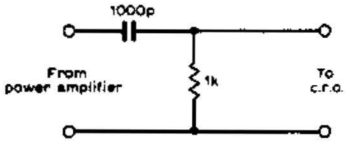


Fig. 6 *Simple differentiator circuit used in tests. The output is 1V for an input rate of change of 1V/μs.*

The experimental procedure adopted was as follows. First a frequency test record was used to check that the system had a flat frequency response, within ±1dB, up to 12kHz. Then a suitable music record was selected, and the system gain was adjusted so that the input to the Fig. 6 circuit occasionally reached peak values of ±10V, but not more. The c.r.o. was then transferred to the differentiator output, the record replayed, and the maximum output voltage excursion from the differentiator during the replay was determined. The test was done with a wide variety of records, including one of the Sheffield direct-cut discs. The largest instantaneous outputs from the differentiator were caused by occasional dust clicks, and went up to over 0.40V, but on the music they never exceeded about 0.14V. The latter corresponds to a slew-rate of 0.14V/μs, which is the peak instantaneous slew-rate of a sine-wave with amplitude ±10V and frequency approximately 2.2kHz.

The implication of the above is that an amplifier with f_{crit} = 2.2kHz, i.e. capable of giving full output on sine-waves up to 2.2kHz, without suffering from slew-rate limitation, and with sufficient freedom from ordinary non-linearity distortion, will reproduce such records entirely satisfactorily. I can almost hear some readers saying "this is ridiculous – it's well established that amplifiers must be free from slew-rate limiting, at full output level, up to at least 20kHz"! But has this, or anything approaching it, in fact, been properly established? I do not think so. But because of such doubts, it is worth approaching the matter from a different angle, as follows.

The maximum instantaneous recorded velocities on records occur over the band extending from about 700Hz to, perhaps, 8kHz, and are normally in the region of 30cm/s[6]. Suppose the gain of an RIAA equalized replay system is adjusted so that a 1kHz sine-wave recording with 30cm/s peak instantaneous velocity gives an output voltage of 10V peak. Since for a sine-wave voltage with peak value $\hat{V}$, the peak rate of change of voltage is $\hat{V} \times 2\pi f$, the peak rate of change of voltage for a 1kHz sine-wave of peak value 10V is 0.063V/μs. It is probably fairly unusual for a peak velocity of 30cm/s to be recorded at a frequency as high as 8kHz, but if this did happen, then, ignoring for the moment the effect of the RIAA equalization, the output slew rate would be 8 × 0.063, i.e. 0.50V/μs. However, at 8kHz, the RIAA equalization introduces a loss of 11.7dB (×3.85) relative to the response at 1kHz, so the figure of 0.50V/μs is reduced to approximately 0.13V/μs. This, it will be seen, ties up surprisingly well with the experimentally determined figure, mentioned above, of 0.14V/μs.

The Fig. 6 differentiator was also used with a master tape recording of violin music with piano accompaniment, thought to be of unusually good fidelity. When adjusted to give a peak replay voltage of 10V as before, the peak instantaneous differentiator output voltage observed was 0.083V, so that the peak slew-rate was 0.083V/μs. A 10V peak sine-wave of 1.3kHz has this same slew rate.

Similar tests done with programme from an f.m. tuner yielded generally equivalent results as far as the actual audio waveform was concerned, but with the complication that, on stereo transmissions, owing to imperfect filtering in the tuner, the (L-R) sidebands greatly increased the peak dv/dt value at the differentiator output, a figure of about 0.4V/μs being obtained with the audio level at ±10V as before. By using the 10kHz filter in the audio control unit, the f.m. multiplex waveform was almost eliminated, the peak slew-rate of the remaining audio waveform being about 0.15V/μs. It is clear that without the filter, the minimum acceptable slew-rate limit in the audio amplifier would be determined largely by the amount of f.m. multiplex waveform present in the tuner output, since unpleasant intermodulation effects can occur if the amplifier is unable properly to follow this waveform. The amount of multiplex waveform in the output of f.m. tuners varies a great deal from one make to another.

The above quite low slew-rates will seem less surprising when it is remembered that the success of the pre-emphasis and de-emphasis schemes universally used in both recording and f.m. broadcasting systems is dependent largely on the fact that the high-frequency components of all normal audio waveforms are of much smaller amplitude than the lower frequency components.

Necessary amplifier slew-rate limit

Provided an amplifier is not overloaded, and provided it has sufficient feedback to make the distortion when not slew-rate limiting adequately low, there is certainly no absolute necessity for the slew-rate limit of the amplifier to be any larger than the maximum rate of change, or slew-rate, of the waveforms handled by it. This point needs emphasising, for reading Jung's interesting articles can easily make one jump to the conclusion that there is a *fundamental* need for the amplifier slew-rate limit to exceed the maximum rate of change of the programme waveform by a large factor. That this cannot possibly be true may be seen by imagining, or actually making, an amplifier with the same broad configuration as in Fig. 1, but in which Tr_1 is replaced not by a simple long-tailed-pair, but by a more complex circuit having a large amount of internal feedback. Then the distortion of the part of the amplifier that precedes C will remain extremely low right up to the slew-rate-limit overload point. Such an amplifier will fail to satisfy Jung's "new slew-rate criterion" by a very large factor, and yet, provided the distortion in the output stage etc. is sufficiently low, it will give no subjectively detectable quality degradation on any normal programme material.

With an ordinary long-tailed-pair input stage, the distortion introduced by it will be mainly third-harmonic, with the higher-order harmonics well subdued, provided the amplifier slew-rate limit is made higher than the maximum slew-rate of the programme by a reasonable factor, say two or three times. The distortion will then be of much the same character as that introduced by a good tape recorder, but will be of appreciable magnitude only at high audio frequencies. Provided the distortion is held down to a reasonably low magnitude – well under that of a recording system, to be on the safe side – by sufficient overall feedback, it will not be subjectively detectable. □

References

1. Jung, W. G., Stephens, M. L. and Todd, C. C., "Slewing induced distortion in audio amplifiers", Feb. 1977 articles series preprint, *The Audio Amateur*, Box 176, Peterborough, New Hampshire 03458 (USA).
2. Jung, W. G. Stephens, M. L. and Todd, C. C., "Slewing induced distortion and its effect on audio amplifier performance – with correlated measurement/listening results," AES Preprint 1252, AES Convention May 1977.
3. Jung, W. G., "Slewing induced distortion," *Hi-Fi News*, Nov. 1977, pp.115-123.
4. Baxandall, P. J., "Low distortion amplifiers – Part 2," *J. British Sound Recording Association*, Nov. 1961, pp.246-256.
5. Taylor, E. F., "Distortion in low-noise amplifiers," *Wireless World*, August 1977, pp.28-32.
6. Kogen, J. H., "Gramophone-record reproduction: development, performance and potential of the stereophonic pickup," *Proc. IEE*, vol. 116, No. 8, August 1969, pp.1338-1344.

Audio power amplifier design — 2

Negative-feedback concepts

The best result of mathematics is to be able to do without it — OLIVER HEAVISIDE

by **Peter J. Baxandall,** B.Sc.(Eng), F.I.E.E., F.I.E.R.E.

In the January issue the concept, and possible consequences, of slew-rate limitation were discussed, with particular reference to one cause, in which the first stage of an amplifier is unable to supply the current demanded by the collector-to-base feedback-stabilization capacitor in the second stage. With suitably modified circuit designs such effects may be made insignificant. Before specific circuits are discussed in later articles, the present article will deal with some basic ideas about negative feedback and transfer functions.

Feedback terms: definitions

Fig. 1 represents the general case of an amplifier with overall feedback. The + and — signs against the symbols for voltages indicate the polarities that exist when the instantaneous values are called positive. V_{out}/V_{in} is the gain with feedback, or closed-loop gain. A is the forward gain, or open-loop gain. From the diagram it is evident that

$$(\beta V_{out}+V_{in})A=V_{out}$$

(Except at middle frequencies, the + sign must be taken to mean addition taking account of phase angle.)

From the above

$$V_{out}(1-A\beta)=AV_{in}$$

$$\text{or } V_{out}/V_{in}=\frac{A}{1-A\beta} \qquad (1)$$

This formula may be regarded as the universal feedback formula, and is just as relevant to positive-feedback applications such as Q-multipliers and some active filters as it is to negative-feedback amplifiers. At medium frequencies, where it will be assumed there are no unwanted phase shifts, A should be taken as a simple negative number if the amplifier phase inverts, β should be taken as negative if the output from the β network is subtracted from V_{in} instead of being added as shown. For a negative-feedback amplifier $A\beta$ will be negative at medium frequencies.

Sometimes the denominator of (1) is given as $1+A\beta$, and then only the magnitudes and not the signs of A and β are to be inserted in the formula. The formula is specifically a *negative*-feedback formula, and the corresponding formula for *positive* feedback then has a denominator $1-A\beta$. This is surely an unnecessary complication, which can lead to confusion in some applications where it is not immediately obvious whether the feedback is to be treated as positive or negative.

The loop gain is the gain right round the feedback loop, and is $A\beta$ in Fig. 1. This concept is simple enough in the ideal context of Fig. 1, but in many practical circuits some care must be taken when calculating or measuring the loop gain. For example, how do we calculate the loop gain in Fig. 2? If the loop is broken by removing the connection between P and Q, and a test voltage V_t is applied between P earth, then this would produce, at the junction of R_2 and R_3, with Tr_1 removed, a voltage of $V_t\beta$. This voltage is effectively applied to the emitter of Tr_1 in series with a resistance of $R_2R_3/(R_2+R_3)$, which appears in series with $1/g_m$, reducing the effective mutual conductance of the stage. Alternatively we may calculate the value of R_2 and $1/g_m$ in parallel, and use this value in place of R_2 for calculating the actual feedback voltage appearing at the emitter due to the test voltage V_t. In obtaining the relevant output voltage from Tr_2, knowing its collector current, it is necessary to add a load resistor between Q and earth of the same value as that previously provided by the feedback network.

Fig. 3 illustrates the meaning of the terms series, shunt, current and voltage feedback. It will be seen that the convention is that 'series' and 'shunt' relate to the way the feedback is injected into the input circuit, whereas 'voltage' and 'current' relate to the manner in which the feedback is derived in the output circuit. Voltage feedback causes the load to be fed as from a generator whose internal impedance, or output impedance as it is often called, tends to zero as the amount of feedback is increased, whereas current feedback causes the output impedance to tend to infinity with increasing feedback.

Fig. 4 shows how a combination of voltage and current negative feedback may be used to produce an amplifier with a prescribed value of resistive output impedance, such as might be required, for example, when feeding into a telephone line. This technique is less wasteful of available output power capability than is the alternative of

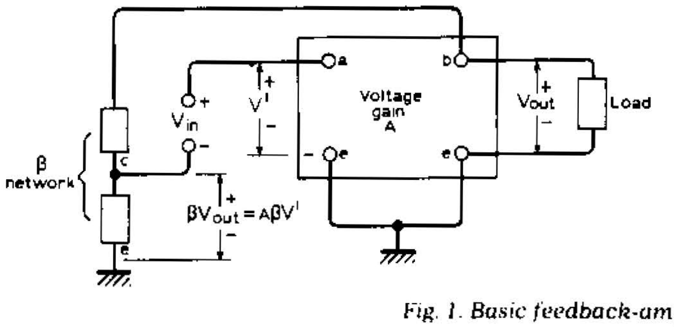
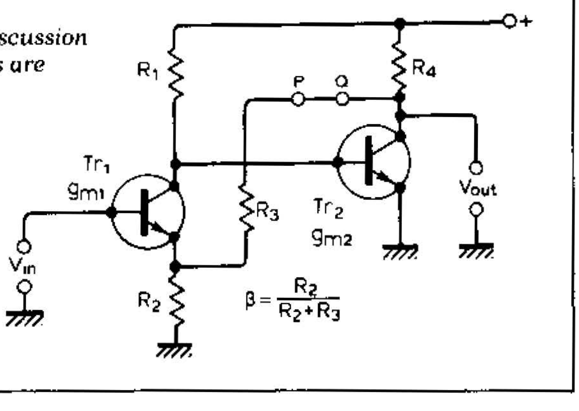


Fig. 1. Basic feedback-amplifier circuit.

Fig. 2. Circuit relating to discussion of loop gain. Biasing details are omitted for clarity.



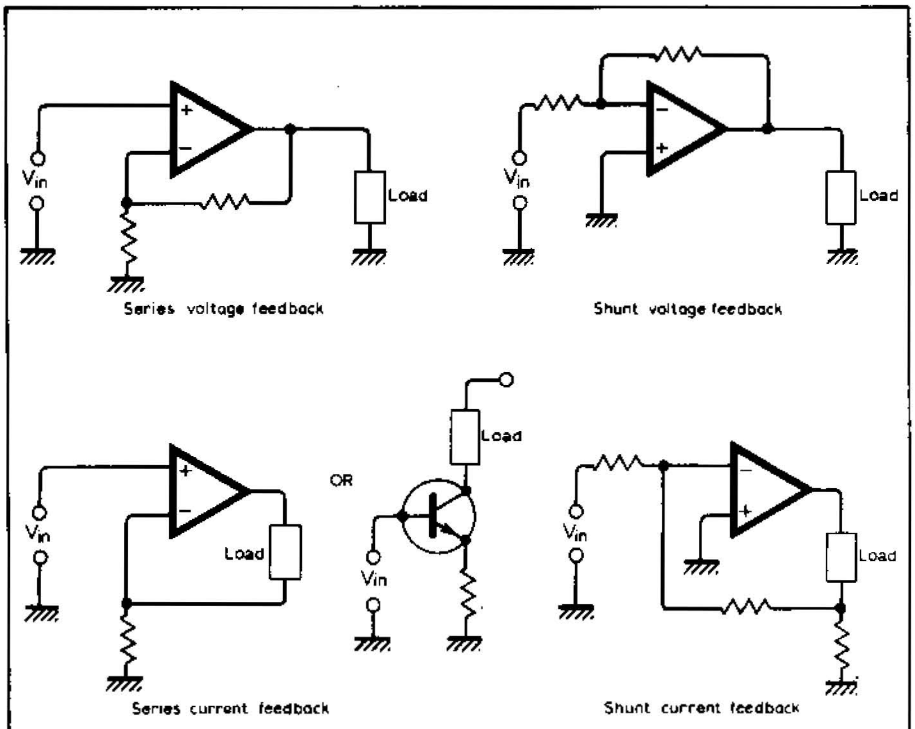


Fig. 3. Four different types of negative feedback.

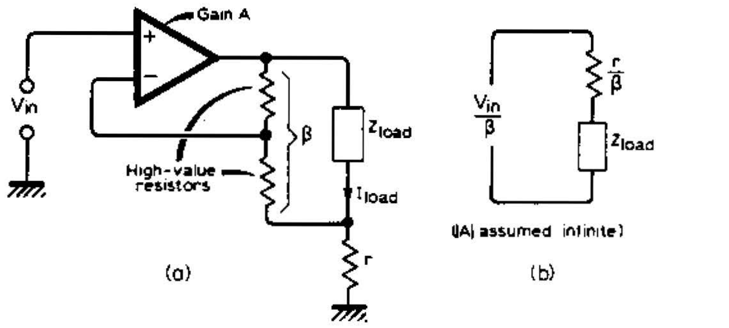


Fig. 4. (a) Feedback circuit with combined voltage and current feedback; (b) equivalent circuit as seen by load.

using an amplifier with simple voltage or current feedback, in association with a resistor equal in value to the required output impedance.

Considering Fig. 4(a), and assuming the ideal case of an infinite-gain amplifier, it is evident that

$$\beta V_{load} + r\, I_{load} = V_{in}$$

or $\beta(Z_{load}\, I_{load}) + r\, I_{load} = V_{in}$

which gives

$$I_{load} = \frac{V_{in}}{r+\beta Z_{load}}$$

or

$$I_{load} = \frac{V_{in}/\beta}{r/\beta + Z_{load}} \qquad (2)$$

This shows that the equivalent circuit must be as in Fig. 4(b). By arranging for the voltage drop across r to provide positive instead of negative feedback, a negative resistive output impedance may be obtained.

Amplifiers are often said to have x decibels of negative feedback at a specified frequency, and such a statement is open to more than one possible interpretation. It is sometimes taken to mean that $20\log_{10}|\text{loop gain}| = x$, but the normal and preferred meaning is that the amount of negative feedback is such as to reduce the amplifier gain by x dB, due precautions being taken to maintain equal loading conditions before and after closing the loop, as already explained. A little thought in relation to equation (1) will show that these two definitions of the amount of negative feedback are not precisely equivalent, and differ quite significantly when the amount of feedback is small. With the preferred definition, feedback is negative at a given frequency if it reduces the gain and positive if it increases the gain. Frequently a practical negative-feedback amplifier will exhibit a peak in its frequency response at high frequencies, near the unity-loop-gain frequency. In the region of the peak, the gain may be higher with feedback on than without it, so that the intended negative feedback has here become positive feedback.

It is sometimes said that feedback is negative if the real component of the feedback voltage, βV_{out}, is in antiphase with V'_{in} Fig. 1, V' itself being taken as purely real, and that feedback is positive if the real component of βV_{out} is in phase with V'. This, however, is a popular misconception, and is quite inconsistent with the distinction between positive and negative feedback given above – as will become evident from the discussion of phase relationships later in this series.

Stability considerations

The subject of stability in feedback systems is a vast one, on which many learned and highly mathematical treatises have been written. The most famous are probably those of H. Nyquist[1] and H. W. Bode[2], both of Bell Telephone Laboratories. Though old, these contributions deal with the fundamentals of the subject thoroughly and in depth, and are still regarded as absolutely sound. Many electronic engineers such as myself, particularly those lacking any formal training in feedback theory, are liable to feel rather overwhelmed by the amount and complexity of the available literature, and concepts such as complex frequency, poles and zeros, contour integration, the Heaviside operator, Laplace transforms and signal-flow graphs seem like insurmountable barriers to some people. However, I believe that the vital thing is to acquire sufficient theoretical understanding to be able to appreciate vividly the reasons for the various effects that occur, and what the available possibilities are for modifying the circuit design as first conceived to give optimum performance. The amount of detailed theoretical background necessary to achieve this is in fact surprisingly small – though some of the mathematical enthusiasts will probably deny this!

There are several reasons why it is unnecessary for a good amplifier designer to know as much mathematical feedback theory as is sometimes supposed. Firstly, much of the fundamental analysis was originally done to *find out* what the stability criteria were, and how they could be expressed in forms convenient for engineers to use. This having been done, and being well established, the engineer can use the results without needing to be able to prove them. Secondly, provided there is a proper qualitative understanding of the problem, the precise optimum values of some components are often best determined experimentally. This is largely because, at the quite high frequencies involved – which may extend up to several MHz – some degree of approximation to the true transistor behaviour would inevitably have to be adopted in a purely theoretical, perhaps computer-aided, design approach. Some people may say that arriving at optimum values for some components by trial and error does not constitute a respectable modern design technique, but I cannot agree with this outlook. One way to regard such a trial-and-error approach is to say that one is using the actual amplifier circuit itself as an analogue computer – changes are made to the circuit values and the results are displayed in analogue form on an oscilloscope. If carried out in an intelligent manner, this seems to me to be a much more direct, economical and generally sensible technique than that of forming a mathematical model of the circuit for processing by a digital computer, but I recognise that what is best done depends a good deal on the background and preferences of the designer.

In some quarters there is a belief that the circuit designer himself should spend his time in an office with paper and a computer, and leave the practical work to others, but I do not think that this philosophy is the most effective one. Experimental work is very stimulating – some unexpected effect is observed, and in a flash one may see that a modification to the circuit would be an improvement. This can often be tried immediately, and may lead to prolonged thought and further ideas. At some point a theoretical analysis may be called for, followed by more experimenting. It is this continuous alternation of experimental and theoretical activity that leads, in my experience, to the evolution of novel and improved designs. Of course, an almost inevitable result of such activity is often that what started off as a neat experimental board tends to have become a somewhat untidy bird's nest at a later stage. However, I think most amplifiers having any real originality of design have probably evolved through such a stage before reaching that of an elegant printed-circuit board.

A very real danger is that if an engineer becomes too absorbed in advanced mathematical techniques, he may fail to give enough attention to other more down-to-earth, but very important, aspects of the overall design work. In a contribution some years ago[3], I said "whilst it is virtuous to be able to analyse a circuit, it may be even be more virtuous to be able to see that a detailed analysis is unnecessary, or to invent a better circuit whose behaviour is more easily predicted."

The aim in what follows will be to present the minimum theoretical background which is thought to be necessary for anyone undertaking to design the feedback stabilization aspects of an audio amplifier with understanding and in a properly optimized manner. Little more than the j-notation[4] will be employed. However, some readers will doubtless wish for a rather broader background of theory, since much published literature on amplifier design uses the concepts of complex frequency, poles and zeros etc. At a fairly elementary level, the excellent series of articles by "Cathode Ray" (M. G. Scroggie) in this journal in 1962 may be recommended[5, 6, 7, 8]. A more advanced and complete treatment of feedback theory and practice will be found in a very good book "Amplifying Devices and Low-Pass Amplifier Design" by Cherry and Hooper[9]. Though they do not hesitate to use determinants etc. when thought to be appropriate, a true engineering outlook is evident and the book contains much very enlightened practical advice on design aspects.

In a.c. coupled amplifiers, stability problems arise at both low and high frequencies. Only the high-frequency problems will be considered here, i.e. all circuits will be treated as d.c. coupled amplifiers, but the principles discussed are very easily adapted, in common-sense ways, to the low-frequency situation when necessary.

Some simple notions about transfer functions will first be considered, because understanding these helps one to appreciate better how the whole negative-feedback story fits together. A transfer function for a feedback amplifier, or any other circuit, is simply an equation giving V_{out} as a function of V_{in}. It is normally assumed that the amplifier is free from non-linearity distortion, but apart from this reservation, the transfer function contains all the necessary information about the frequency response, phase response, transient response and stability margins of the amplifier. The snag is that, except in quite simple cases, deriving and simplifying the transfer function for a feedback amplifier is exasperatingly tedious, even for those with a natural aptitude for such things, which I certainly do not have! The Nyquist diagram, and Bode amplitude and phase plots considered later, represent a vastly more convenient and practicable approach for most amplifier design purposes.

However, it is always theoretically possible simply to use the j-notation to calculate the currents and voltages everywhere in the amplifier circuit due to V_{in} and V_{out}, and thus to form the transfer-function equation. Purely as an illustration of the ideas involved, consider the simple and somewhat idealized circuit of Fig. 5. Using the j-notation gives the current in C_2 as $j\omega V_{out}C_2$. The current in R_4 in the direction shown is V_{out}/R_4. The current in R_3 is the sum of these currents, enabling one to calculate V'. Continuing on these lines leads to the result:

$$V_{in} = -V_{out}R_{in}/R_1[1 + j\omega C_2R_3 + R_3/R_4 + j\omega C_1R_2\ (1 + j\omega C_2R_3 + R_3/R_4) + j\omega C_2R_2 + R_2/R_4 + R_1/R_4] \qquad (3)$$

This as it stands is not much use, for one cannot easily see the physical significance of it. The vital thing when deriving transfer functions is to continue until they have been got into a nice tidy, recognisable form. By collecting terms and rearranging, equation (3) can be got into the form:

$$V_{out}/V_{in} = K \times \frac{1}{1 + j\omega T_1 - \omega^2 T_2{}^2} \qquad (4)$$

K in this is given by:

$$K = \frac{R_1R_4}{R_{in}(R_1 + R_2 + R_3 + R_4)} \qquad (5)$$

T_1 and T_2 are time constants, each given by a somewhat cumbersome expression with several terms in. One can, moreover, very usefully go a stage further than (4), and get it into the form:

$$V_{out}/V_{in} = K \times \frac{1}{1 + (1/Q)j\omega T - \omega^2 T^2} \qquad (6)$$

Here T is obviously equal to T_2 of equation (4), and we also must have $(1/Q)T = T_1$, giving $Q = T/T_1$, i.e.:

$$Q = T_2/T_1 \qquad (7)$$

Now the physical significance of (6) is instantly apparent if one knows how to "read" it. Q is the Q of a tuned circuit arranged as in Fig. 6(a), having a resonance frequency given by $\omega_0 = 1/T$.

Sometimes transfer functions such as

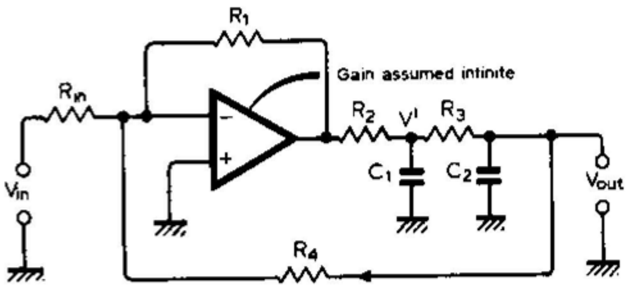


Fig. 5. Circuit relating to discussion of transfer functions.

(6) are given in the form:

$$V_{out}/V_{in} = K \times \frac{1}{1 + (1/Q)pT + p^2T^2} \quad (8)$$

Comparing (6) and (8) it is evident that p = jω. Though it is perfectly all right, in a sine-wave context, to regard p simply as a convenient abbreviation for jω, its full significance is much deeper, for it is Heaviside's operator and means d/dt. Equations such as (8) are thus applicable not only under sine-wave conditions, but also for any other kind of input waveform. Mathematical techniques are available whereby, given the amplifier transfer function, the output waveform resulting from a voltage step or other transient input may be calculated. But in view of the ease with which such responses may be obtained using an oscilloscope, the actual need for such mathematical techniques seldom if ever arises in normal amplifier design work, in my experience. Sometimes when the transient response of an experimental amplifier circuit is under consideration, it is convenient to make up a little simulator circuit, in which all time-constants have been increased by a factor of, say, a thousand compared with the real circuit. The idealized response can thus be obtained, and the relationship between this and the response of the original circuit may shed light on the significance of stray capacitance or other overlooked effects in the latter. The ready availability of type 741 operational amplifiers makes it very quick and easy to do such tests.

Heaviside's operational calculus tends to be somewhat out of favour nowadays, but a very strong case in its favour is presented by two authors from the BBC Research Department in reference 10. It is argued that the technique gives a much better physical insight into the nature of the problem being investigated than do the alternative mathematical techniques available.

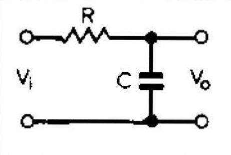
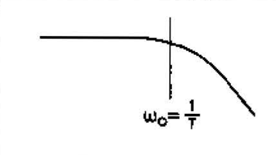
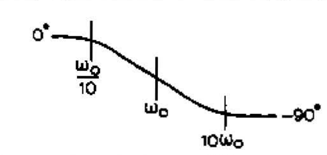
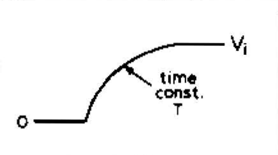
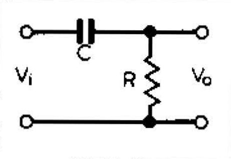
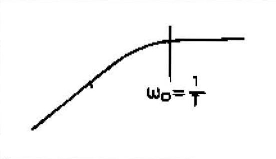
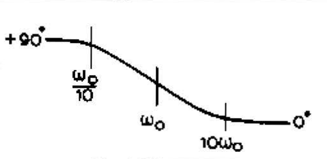
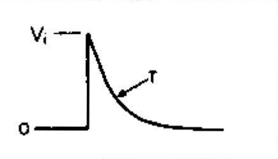
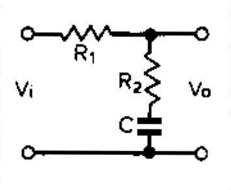
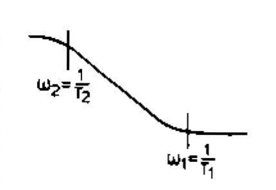
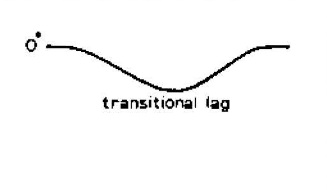
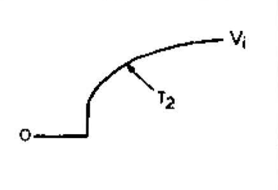
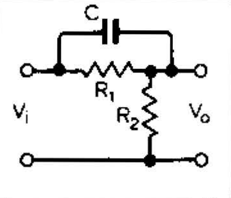
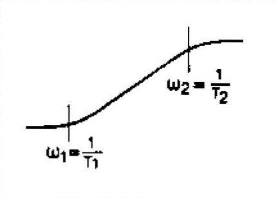
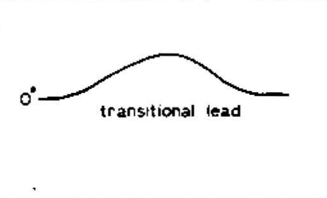
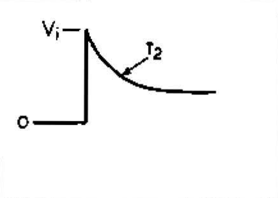
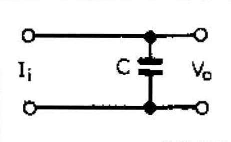
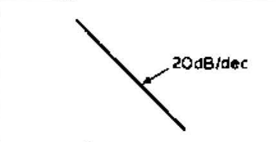
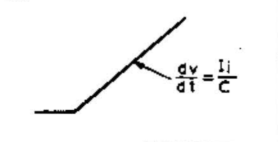
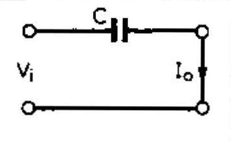
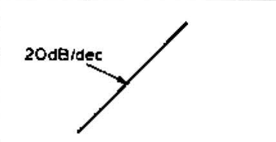
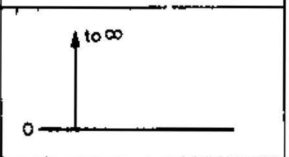
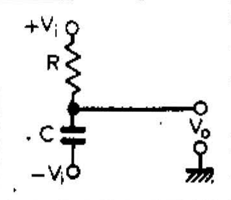
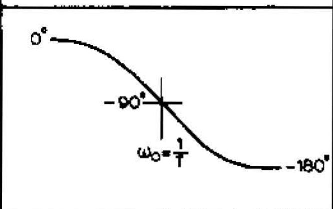
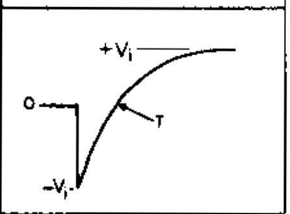
For amplifier designers, the important things to appreciate about transfer functions may be summarized as follows:

(a) Any linear network or amplifier has a transfer function.

(b) However complex the network or amplifier may be, the denominator of the transfer function — if you're clever enough — can be got into the form of a number of factors, which may be either quadratic ones as in equation (8), or simpler ones of the form (1 + pT).

(c) If any of the quadratic factors in the denominator have negative Q, i.e. negative damping, the system will be unstable.

(d) The numerator can take various forms according to whether the system has a low-pass, band-pass or high-pass type of response, and whether there are notches in the frequency response or not.

	Circuit	Transfer function	Frequency response	Phase response	Step response
A		$\frac{V_o}{V_i} = \frac{1}{1+pT}$ $T = CR$			
B		$\frac{V_o}{V_i} = \frac{pT}{1+pT}$ $T = CR$			
C		$\frac{V_o}{V_i} = \frac{1+pT_1}{1+pT_2}$ $T_1 = CR_2$ $T_2 = C(R_1+R_2)$			
D		$\frac{V_o}{V_i} = K \times \frac{1+pT_1}{1+pT_2}$ $T_1 = CR_1$ $T_2 = C \times \frac{R_1R_2}{R_1+R_2}$ $K = \frac{R_2}{R_1+R_2}$			
E		$\frac{V_o}{I_i} = \frac{1}{pC}$		−90°	
F		$\frac{I_o}{V_i} = pC$		+90°	
G		$\frac{V_o}{V_i} = \frac{1-pT}{1+pT}$ $T = CR$	flat		

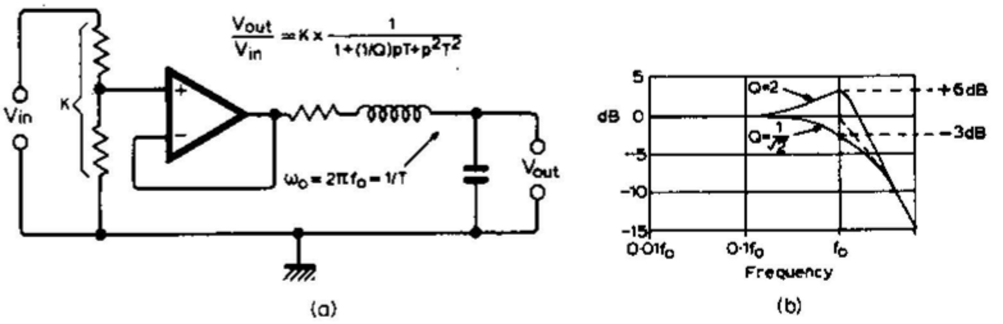

$\frac{V_{out}}{V_{in}} = K \times \frac{1}{1+(1/Q)pT+p^2T^2}$
Vin
K
+
−
$\omega_0 = 2\pi f_0 = 1/T$
Vout
(a)
5
dB 0
−5
−10
−15
Q=2
$Q=\frac{1}{\sqrt{2}}$
+6dB
−3dB
0·01f₀
0·1f₀
f₀
Frequency
(b)

Fig. 6.(a) Circuit giving same response as Fig. 5; (b) and (c) show the frequency response and the step response respectively for two values of Q. $Q=1/\sqrt{2}$ gives second order Butterworth response.

(e) Any required response characteristic whatever can be obtained from a combination of suitably-designed feedback amplifiers, without the need for any inductors, this being the basis of the whole subject of active filters.[11]

Though it is seldom sensible to try to derive the overall transfer function of a complete feedback amplifier, except in the relatively simple cases which usually apply in active-filter design, it is quite important to be able to derive the transfer functions of parts of the circuit of a feedback amplifier, for this is really the basis of most practical design work on such amplifiers. The table gives some simple networks familiar to most readers, together with their transfer functions and frequency, phase and step-input responses. The relevance of the all-pass case G will become evident later. Though the transfer functions may be worked out using the j-notation, and p substituted for jω at the end, it is really more convenient to work with p from the beginning. Thus the impedance of a capacitor is $1/pC$ and the impedance of an inductor is pL. Suppose, for example, we have R and C in parallel. The total impedance is given by

$$Z = \frac{R \times (1/pC)}{R + (1/pC)}$$

Multiplying top and bottom by pC gives

$$Z = \frac{R}{1+pCR} \qquad (9)$$

This is therefore the ratio V_{out}/I_{in} for the network, and as would be expected it has the same form of transfer function as network A in the table.

A simple illustration of the practical utility of thinking of transfer functions in terms of p rather than jω arises if one considers the problem of determining the output waveform to be expected from network B in the table when the input waveform is a linear voltage sweep, or ramp. One simply "operates upon" the input waveform with bits of the transfer function in turn, chosen in the order that makes things easiest. Thus the ramp waveform multiplied by pT, i.e. differentiated, gives a step waveform. The step multiplied by $1/(1+pT)$ gives an exponential output waveform as shown at the top right-hand corner of the table. A particularly lucid and easy-to-understand paper dealing with topics such as this was written just after the war by Professor F. C. Williams[12]. Though the practical circuits are, of course, all valve ones, the lengthy discussion of the overall design philosophy is highly relevant to present-day problems. The aim was to evolve reliable circuits of precision performance, suitable for trouble-free production, using the minimum of mathematics. Acknowledgement is made to A. D. Blumlein for having provided much of the early inspiration for this work. Some of these pulse circuit ideas are of greater interest to audio engineers than in the past, even in the non-digital field, because of the increased attention now being given to transient response and impulse measuring techniques.

In planning the feedback stabilization details for most audio amplifiers, the normal practice is to think in terms of the rate at which the loop gain is attenuated with rising frequency, bearing in mind all along that the transient behaviour is closely related to this. The relevant techniques will be discussed in the next article.

Corrections to January 1978 article

In Fig. 1, a resistor should be inserted in series with Tr₄ emitter. The arrow in Tr₁ collector lead should be labelled "I_{dc}." In equation (6), the denominator should be "$2\pi\hat{V}_{in}$". The equation just below equation (6) is completely wrong and should be:

$$\frac{\text{slew-rate limit}}{\hat{V}} = 2\pi f_{crit} \qquad (7)$$

In Fig. 3(a), the top waveform was inadvertently cut off at the bottom and should be a complete sinewave. Apologies for the bad reproduction of these waveforms. In the fourth line of the footnote on page 55, the word "is" should be inserted before "approximately". On page 56, first column, 14 lines from the bottom, the word "amplifier" should be inserted between "the" and "slew-rate".

References

1. Nyquist, H., Regeneration Theory, *Bell System Tech. J.*, Jan. 1932, p.126.
2. Bode, H. W., Network Analysis and Feedback Amplifier Design. (van Nostrand 1945).
3. Baxandall, P. J., Papers for the Practising Designer, Letter to Editor, *J.I.E.R.E.*, Dec. 1968.
4. Cathode Ray, "j", *Wireless World*, Feb. 1948.
5. Cathode Ray, Transfer Functions, *Wireless World*, April 1962, pp.177-181.
6. Cathode Ray, Poles and Zeros, *Wireless World*, May 1962, pp.225-229 and June 1962, pp.289-294.
7. Cathode Ray, Differential Equations, *Wireless World*, July 1962, pp.333-337.
8. Cathode Ray, Excitations and Responses, *Wireless World*, Aug. 1962, pp.379-383, Sept. 1962, pp.447-450 and Oct. 1962, pp.507-511.
9. Cherry, E. M. and Hooper, D. E., Amplifying Devices and Low-Pass Amplifier Design. (John Wiley 1968).
10. Head, J. W. and Mayo, C. G., Unified Circuit Theory in Electronics and Engineering Analysis. (Iliffe 1965).
11. Girling, F. E. J. and Good, E. F., Active Filters, *Wireless World*, Aug. 1969 to Dec. 1970 inc. 16 parts; see particularly Sept. 1969, pp.403-408. (Note: In these articles q is used in place of Q in equations such as my eqn. (8), Q being reserved for bandpass filters, where it has a somewhat different significance.)
12. Williams, F. C., Introduction to Circuit Techniques for Radiolocation, J.I.E.E., Vol. 93, Part IIIA, No. 1, pp.289-308 (1946). □

Audio power amplifier design — 3

Nyquist and Bode diagrams

"Design methods suitable for a variety of applications can never be reduced entirely to a set of rules" — H. W. Bode

by **Peter J. Baxandall,** B.Sc.(Eng.), F.I.E.E., F.I.E.R.E.

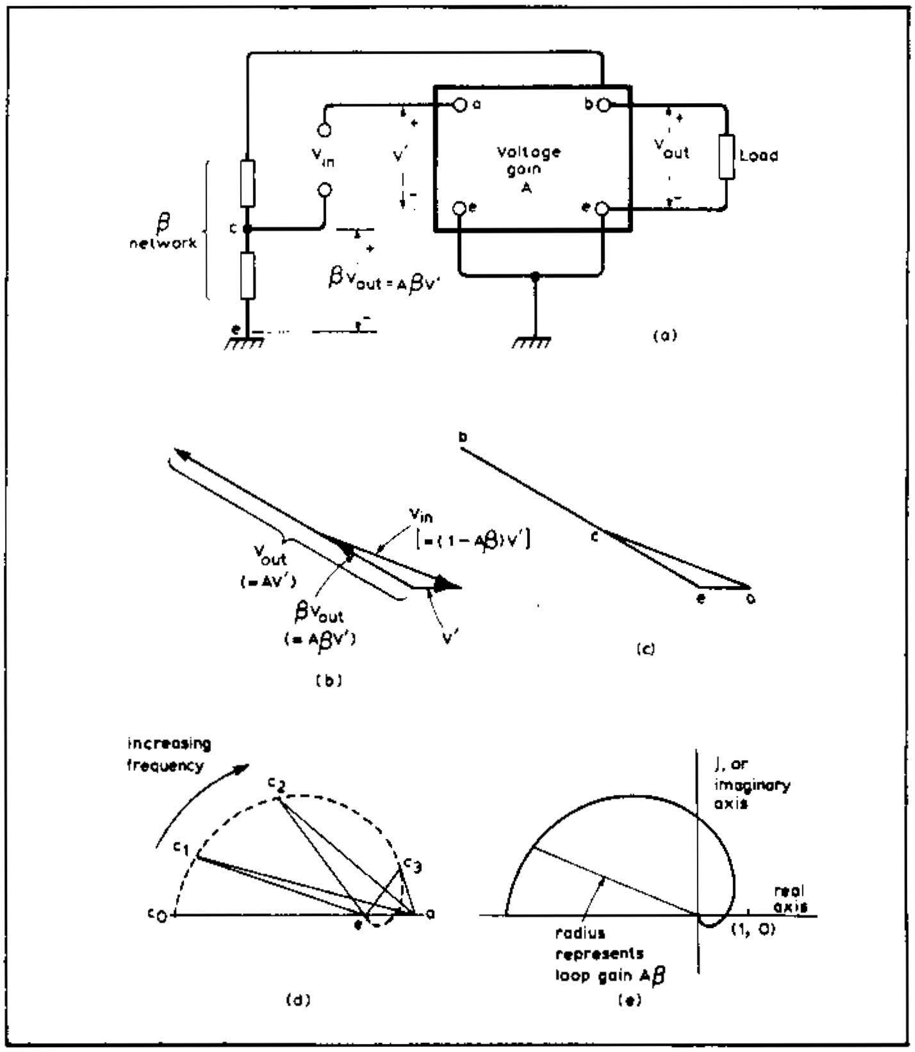
In the March issue it was explained that, assuming negligible non-linearity distortion, the closed-loop transfer function for a feedback amplifier gives full information about the frequency response, phase response, and transient response. In principle, therefore, all theoretical design work could be done by choosing the circuit configuration and values to yield a desired transfer function. However, this is such a tedious and inflexible approach for most amplifier design purposes that other techniques are much preferred.

Although the Nyquist Diagram is seldom actually drawn by amplifier designers, it is the best starting point one can make to gain an understanding of the preferred techniques used in amplifier design. For clarity, Fig. 1(a) has been included even though it is a repeat of Fig. 1 in the March issue. Figure 1(b) shows a phasor diagram for this circuit, drawn in the conventional manner and, for simplicity, the β-network is assumed to give attenuation but no phase shift. Figure 1(c) gives the phasor diagram for the circuit, drawn in accordance with the neater and generally much preferable scheme advocated by M. G. Scroggie[1], in which points on the phasor diagram are lettered to correspond to points on the circuit diagram, neither arrow heads nor voltage symbols then being required. With either scheme, if one likes, the whole phasor diagram may be envisaged as rotating, conventionally anti-clockwise. Then the *vertical* distances between the ends of the phasors represent instantaneous voltage values. Therefore, at the instant of time depicted by the angular position of the diagram shown in Fig. 1(c), points b and c are positive with respect to e. The lengths of the phasors, of course, represent the corresponding peak, or, if preferred, r.m.s. voltage values. The more I use the Scroggie method of drawing phasor diagrams, the more I like it, and my only regret is that, through sheer inertia, I did not change over to it far sooner.

Both of the phasor diagrams shown represent the conditions in the circuit of Fig. 1(a) at one frequency only, and the Nyquist diagram can be regarded as being derived from a set of such phasor diagrams covering all frequencies. These are drawn on the convenient basis that V' has the same value for all of them, being represented by ea in diagram Fig. 1(d). (Usually only the voltage at c is included in the diagram, b is left out.) Therefore, as the frequency is varied, a succession of voltage phasor diagrams, as shown in Fig. 1(d), is obtained. In this example, for simplicity, the amplifier is assumed to be d.c. coupled, so that at zero frequency the feedback voltage βV_{out}, or $A\beta V'$, is exactly in antiphase with the voltage V'. The locus of c, shown as a broken line in Fig. 1(d), is essentially the Nyquist diagram. Normally, however, the quantities plotted in a Nyquist diagram are not voltages, but gains, and are obtained by dividing all the quantities in the phasor diagrams shown in Figs. 1(b), (c) and (d) by V'. The Nyquist diagram in its normal form therefore appears as shown in Fig. 1(e), and is an Argand diagram showing how the loop gain $A\beta$ varies in amplitude and phase as the frequency is varied. Nevertheless, for some purposes, it is more convenient to think in terms of voltage phasor diagrams.

At low frequencies, especially when the loop gain is much larger than depicted here, the feedback voltage βV_{out}, represented, for example, by ec_1, is nearly equal in magnitude to the signal input voltage c_1a, so that the gain of the

Fig. 1. *Basic feedback-amplifier circuit, with voltage-phasor diagrams and Nyquist diagram.*



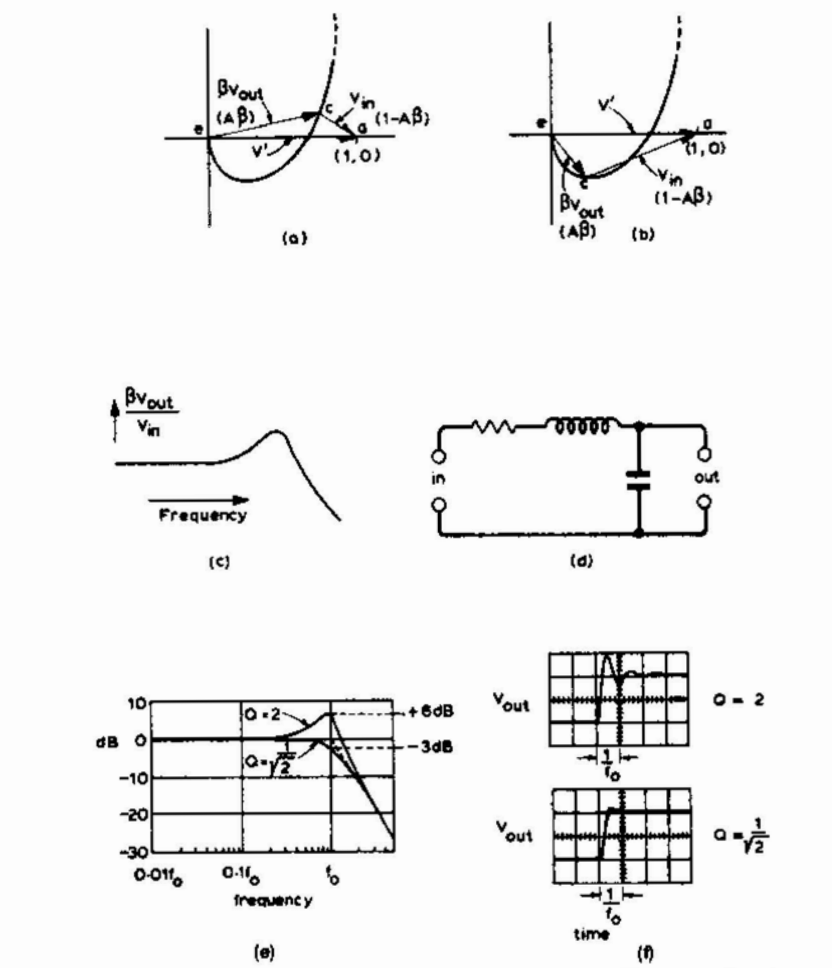


Fig. 2. *(a) and (b) show enlarged Nyquist/voltage-phasor diagrams for the critical region. (c) shows the type of frequency response resulting from (a) and (b), the circuit (d) having an approximately similar sort of response. (e) and (f) show accurate frequency and step responses for the (d) circuit for two values of Q.*

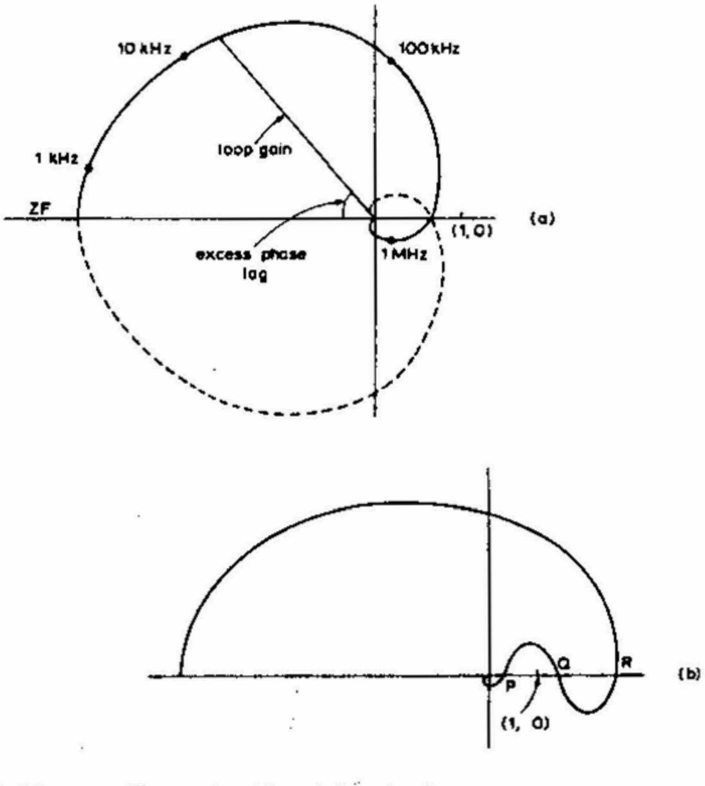


Fig. 3. *Diagrams illustrating Nyquist's criterion.*

amplifier, V_{out}/V_{in}, approximates closely to $1/\beta$*. Consider now the phasor relationship existing at a much higher frequency, when the tip of the $A\beta$ phasor has swung round to the region of the Nyquist diagram (in Fig. 1(e)) which is close to the point (1, 0). The details may be shown more clearly by redrawing just the relevant parts of the diagram to a larger scale. This has been done on a voltage basis in Fig. 2(a) and 2(b), but with the corresponding dimensionless, or gain, quantities shown in brackets. Also, both the conventional and the Scroggie representations have been combined, to suit all readers! In Fig. 2(a) it is seen that the βV_{out} phasor is now much longer than the V_{in} phasor, making the amplifier gain with feedback much greater than $1/\beta$. At even higher frequencies, as shown in Fig. 2(b), βV_{out} has become much less than V_{in}, so that the gain of the complete amplifier is now much less than $1/\beta$. From this it is evident that the closed-loop frequency response will be of the form shown in Fig. 2(c), and that if the Nyquist diagram goes very close to the point (1, 0), the peak in the frequency response will be of large magnitude. To obtain such a frequency response with purely passive elements, an arrangement such as that shown in Fig. 2(d) would be required, and it is obvious that this circuit, if supplied with a voltage step input, will ring if the Q-value is high enough. The frequency and step responses for two values of Q are shown in Figs. 2(e) and 2(f) respectively.

Since the frequency response of an amplifier whose Nyquist diagram passes close to the point (1, 0) is broadly similar to that of a passive circuit such as that shown in Fig. 2(d), it seems reasonable to expect, on these grounds alone, that the amplifier, like the passive circuit, will exhibit very ringy behaviour if the peak in its frequency response is of large magnitude – and this is, indeed, the case.

It is evident from the above simple phasor diagram considerations that if the Nyquist diagram passes through the point (1, 0), the required value of V_{in} for a finite output at that frequency becomes vanishingly small. Oscillation will then occur. A difficult question to answer, however, is whether oscillation can occur under any other conditions. Nyquist, in his famous paper of 1932[2], looked very deeply into this problem and enunciated his stability criterion, which is now universally accepted as being correct.

Nyquist's criterion

Nyquist's criterion states that if a Nyquist diagram, as already described, is plotted for all frequencies from zero to infinity, together with its image in the real axis, as shown in Fig. 3(a), the

* Referring to Fig. 1(b), $V_{out}/V_{in} = AV'/(1-A\beta)V' = (1/\beta)\times A\beta/(1-A\beta)$, showing that the gain becomes approximately $1/\beta$ when $|A\beta|>>1$.

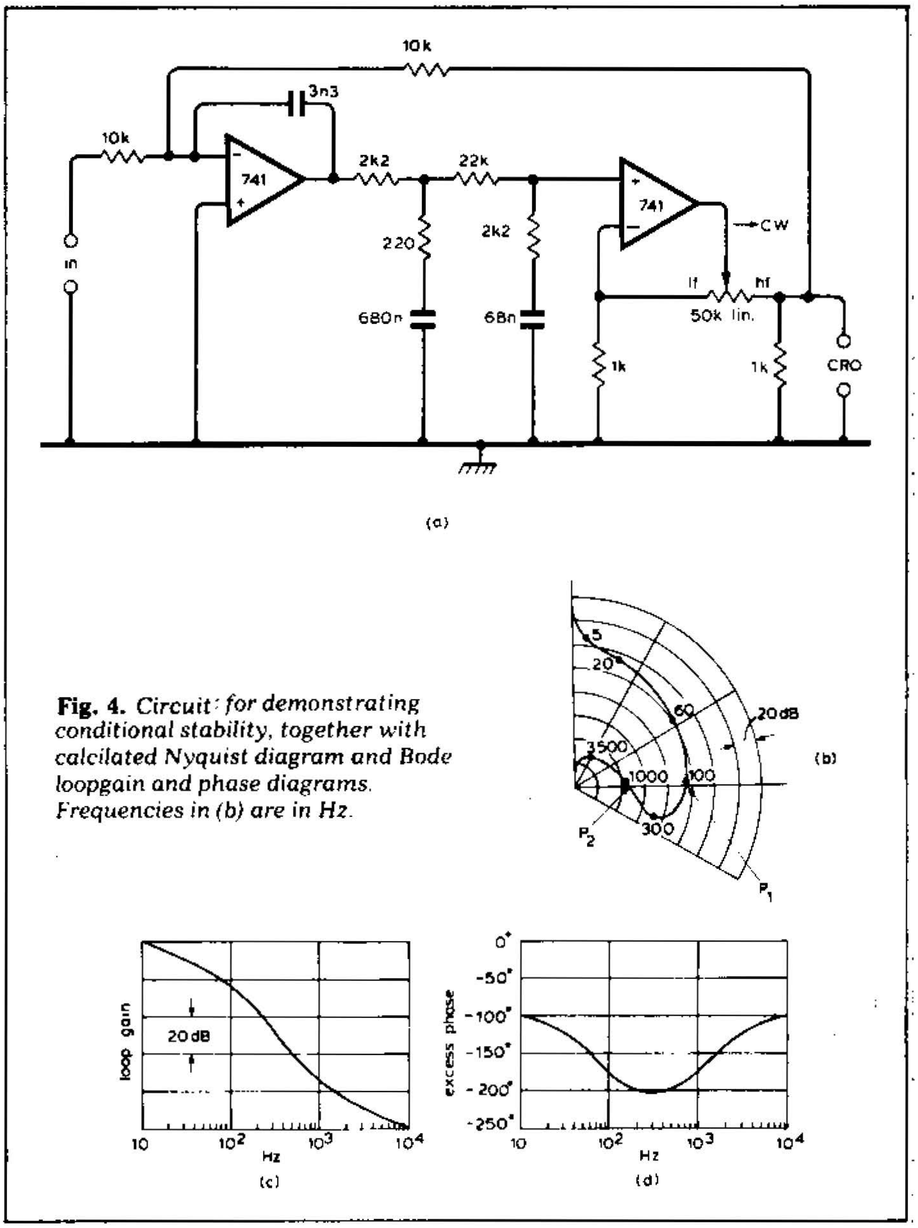


Fig. 4. *Circuit for demonstrating conditional stability, together with calculated Nyquist diagram and Bode loopgain and phase diagrams. Frequencies in (b) are in Hz.*

amplifier will be stable only if the point (1, 0) lies outside the enclosed figure so formed.

The example shown in Fig. 3(a) relates, as before, to a d.c.-coupled amplifier. The angle labelled "excess phase lag" refers to the phase lag that builds up with rising frequency due to shunt capacitances, transistor phase lags etc. The word "excess" is often inserted here to make it clear that the angle referred to does not include the 180° phase angle which is inherent in the fact that the feedback is *negative* at zero frequency. The frequencies marked on the Nyquist diagram are intended to be fairly typical of those which might apply to an audio amplifier. Though the Nyquist criterion, as stated, calls for frequencies from zero to infinity, it is clearly neither feasible nor necessary to cover such a range in practice. It is possible to be caught out, however, if measurements are stopped at too low a frequency, for a Nyquist diagram expected to continue shrinking with rising frequency as it passes the point (1, 0) can occasionally come back away from the origin again in a manner such as to jeopardise stability. This is especially liable to happen when transformers are employed, due to complex resonances involving leakage inductances and winding capacitances.

When complete Nyquist diagrams are plotted, it is usually more convenient to adopt a linear scale of decibels radially, to accommodate the wide range of loop gain magnitudes involved. Sometimes, however, only the part of the diagram fairly near the point (1, 0) need be plotted, and a linear scale may then suffice.

When the loop gain of a feedback amplifier is changed without altering any of the time-constants – for example, by a simple alteration in the value of the overall β – the obvious way to represent this would be to alter the size of the Nyquist diagram, leaving the point (1, 0) fixed. However, a much easier and quicker procedure is to leave the diagram as it is and shift the position of the point (1, 0), effectively altering the scale of the diagram. Usually there is no need to draw the image of the Nyquist diagram in the real axis, as shown in Fig. 3(a), because it is normally obvious whether the point (1, 0) would lie within the complete figure thus formed, without needing to see the broken-line part.

Conditional stability

It is possible to have an amplifier whose Nyquist diagram is something like that shown in Fig. 3(b). With the loop gain adjusted so that the unity-loop-gain point (1, 0) is in the space between P and Q, the amplifier will be stable, for the diagram does not encircle the point. An increase in loop gain, represented by moving the point (1, 0) to the left, will result in the onset of oscillation once the point (1, 0) reaches P. A decrease in loop gain, represented by moving the point (1, 0) to the right, will also cause oscillation (at a different frequency) by the time the point (1, 0) reaches Q. If the reduction in loop gain is sufficient to move the point (1, 0) beyond R, stability will again be achieved.

When operated with the point (1, 0) between P and Q, the amplifier is said to be *conditionally stable*. In this state, it should be noted that, at the frequencies corresponding to the points Q and R, there is zero phase shift round the loop and a loop gain greater than unity – and yet oscillation does not occur. A conditionally stable amplifier may thus be defined as one in which reducing the loop gain causes it to oscillate. An amplifier which remains stable for all values of loop gain setting between the normal one and zero is said to be *absolutely stable*[3]. It is important to distinguish between the term "conditionally stable", as defined above, and a quite different usage of the same term as applied to an amplifier which is stable only if the load impedance satisfies certain conditions. The converse term in this latter case is "unconditionally stable", meaning that the amplifier remains stable with any passive load impedance whatever connected.

Amplifiers having conditional stability in the sense of a reduction in loop gain causing oscillation, are normally to be avoided, and I have never come across a case of one being intentionally used for an engineering application. Their interest lies rather in the light that they throw on one's understanding of the full significance and correctness of Nyquist's criterion.

It is quite easy to make up a simple circuit exhibiting conditional stability, and Fig. 4(a) shows a suitable recipe. This circuit can be made the basis of an excellent and convincing lecture demonstration. With the potentiometer slider at the left, giving low loop gain, the amplifier is absolutely stable, and gives a fairly well damped square-wave response with, say, a 5Hz input square wave. As the slider is moved to the right, increasing the loop gain, the response becomes more and more ringy, going into continuous oscillation at just over 100Hz with sufficient gain increase. Turning the gain up further produces violent oscillation of increasing frequency, as the point (1, 0) is moved across from P_1 to P_2 in the calculated Nyquist diagram shown in Fig. 4(b). As P_2 is reached, the sustained oscillation becomes gentle only, with a frequency of just over 800Hz. Increasing the loop gain still more gives stability once again, but this time it is conditional stability. The more the gain is turned up, the better the damping of the square-wave response becomes, an appropriate square-wave frequency under these conditions being about 100Hz. For demonstration purposes, the output may be reproduced on a loudspeaker. The rather unusual gain control arrangement adopted enables the gain to be adjusted over a very wide range (about 68dB) without the control becoming too "touchy". It is the same arrangement as that used in the BBC Outside-Broadcast amplifier, type OBA/9[4], and such schemes, which combine passive and feedback gain control using a single potentiometer, have many applications. With a c.r.o. fed from the point shown in Fig. 4(a), it is the forward gain of the demonstration circuit that is controlled, the overall β-value remaining constant. Only a small output level can be produced when the potentiometer slider is over to the left, and a c.r.o. sensitivity of 50mV/cm will be found appropriate. An alternative is to feed the c.r.o. from the potentiometer slider, which enables a high output level to be obtained at all potentiometer settings. Now, however, both the forward gain and the overall β-value of the demonstration circuit are being varied, so that the signal gain in the non-oscillating states is dependent on the potentiometer setting.

In the valve era, a strong argument against the use of conditionally stable amplifiers was that the gradual rise in mutual conductances during warm-up caused oscillation to occur before the final conditionally-stable state was reached. As mentioned on page 163 of reference 3, such oscillation, once it started, was liable to persist indefinitely, because of the reduced effective stage gains under overload conditions. It is interesting to note that the demonstration model in Fig. 4 shows no tendencies of this kind if overloaded while in the conditionally-stable state. The only advantage that can be gained by adopting a conditionally-stable design is that it permits a much more rapid attenuation of loop gain with rising frequency than could otherwise be permitted, so that more feedback can be kept in operation up to a higher frequency, with a greater consequent reduction in distortion. Since extremely low distortion can readily be obtained in more straightforward ways, it is probably best to forget about such possibilities.

Gain and phase curves

During design work on feedback amplifiers, most engineers, rather than using Nyquist diagrams, think in terms of curves of gain magnitude and phase, against frequency. The diagrams drawn, which often use straight-line approximations to the true curves, are sometimes called Bode diagrams[5]. Fig. 5 shows simple examples which may be compared with the curves shown in the table in Part 2. Frequently, it is sufficient to draw only the gain diagrams, for provided the networks involved are so-called *minimum-phase-shift networks*, definite relationships exist between the gain and phase characteristics[3]. Then, provided the loop-gain characteristic is designed to meet certain requirements, discussed below, the phase characteristic will automatically be such that stability is assured. In this context, it is obviously necessary to have a very clear conception of just what does, and what does not, constitute a minimum-phase-shift (m.p.s.) network. It has sometimes been said that all the circuits used in ordinary amplifiers are of the m.p.s. type, but this is not necessarily always true. Any circuit in which there is more than one signal path from input to output, is liable not to have m.p.s. characteristics; that is, it is liable to produce more phase shift than necessary for the given gain characteristic. Such a non-m.p.s. network is always equivalent to a m.p.s. network in cascade with an all-pass network, the latter producing phase shift only, without gain variation. A simple example of such a non-m.p.s. circuit, which frequently occurs in amplifiers, is shown in Fig. 6(a). At the frequencies of present interest, the collector resistor R_c exerts negligible shunting effect and may be ignored. At very high frequencies, where C may be regarded as a short-circuit, $V_{out} = I_{in} \times (1/g_m)$.

At lower frequencies the circuit operates as a Blumlein integrator and gives $V_{out} = -I_{in} \times (1/pC)$, where $p = j\omega$. The general relationship is therefore:

$$V_{out} = I_{in}\left(\frac{1}{g_m} - \frac{1}{pC}\right) \qquad (1)$$

or

$$V_{out} = \frac{I_{in}}{g_m}\left(1 - \frac{1}{pT}\right) \qquad (2)$$

where $T = C \times \frac{1}{g_m}$

Now, (2) may be written:

$$V_{out} = -\frac{I_{in}}{g_m} \times \frac{1-pT}{pT}$$

The significance of this may be more vividly seen if it is expressed in the form:

$$V_{out} = -\underbrace{\frac{I_{in}}{g_m} \times \frac{1+pT}{pT}}_{A} \times \underbrace{\frac{1-pT}{1+pT}}_{B} \qquad (3)$$

Part A of Equation 3 represents the response of the network shown in Fig. 6(b), which is relatively innocuous from a feedback-stability point of view. Part B, however, represents an all-pass characteristic (as shown at the bottom of the table on page 44 of the March article), and introduces extra phase lag without affecting the magnitude of the loop gain. Frequently, however, these complications do not significantly affect the stability of an amplifier, be-

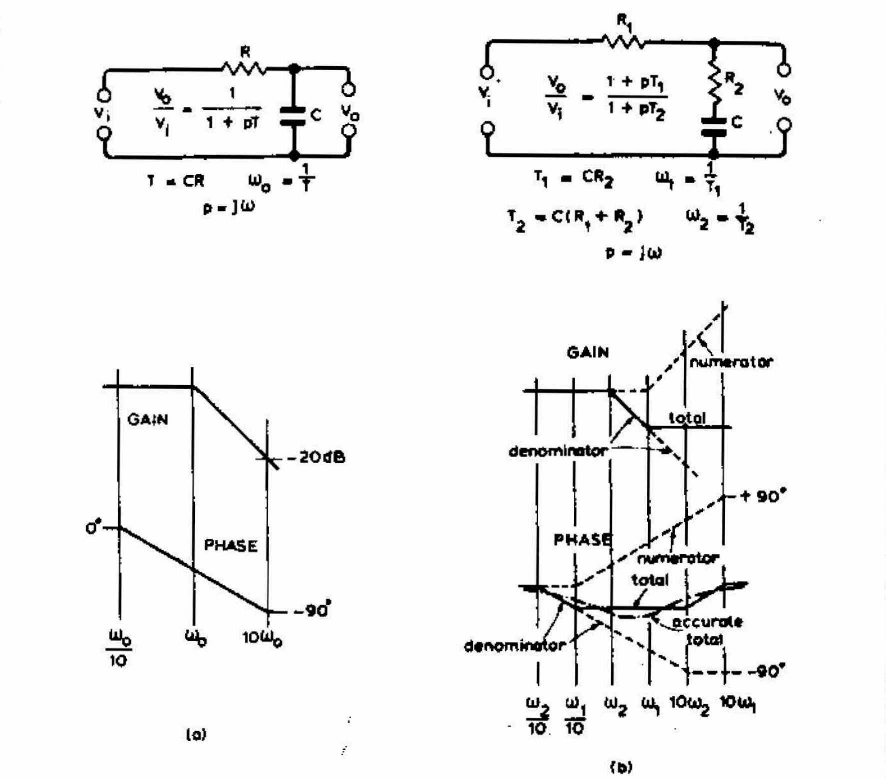


Fig. 5. *Diagrams showing how straightline approximations to the gain and phase characteristics of simple networks may be easily and quickly drawn.*

cause they come in at frequencies well above the unity-loop-gain frequency. For example, with C=100pF and a collector current of 5mA, giving an ideal g_m value of 200mA/V, the frequency at which the all-pass term gives 90° phase lag is, in ideal theory, 320MHz. Sometimes, however, a resistor of, say, 100Ω is included in the transistor emitter lead – maybe as part of a current-limiting scheme – giving much-reduced g_m. This, in association with a higher C value, say 470pF, gives an ideal theoretical all-pass 90° frequency of 3.2MHz, so that the all-pass term will give 10° lag at about 300kHz – not necessarily negligible.

Under overload conditions, the transistor in Fig. 6(a) may be temporarily cut off. Then, the *only* path from input to output is that via C, and the intended phase inversion of the stage is completely lost. With a large amount of overall feedback, which then becomes positive, momentary trigger-action, or high-frequency oscillation, may occur. Unravelling subtle effects such as this – and there are many others – can at times make development work on feedback amplifiers a difficult and demanding exercise.

From the above discussion it will be appreciated that, for most audio amplifier design purposes, it is safe to assume that the networks involved are of the minimum-phase-shift variety, but the possibility of things being otherwise should not be entirely forgotten.

Assuming m.p.s. networks, a high-frequency attenuation rate of 20dB/decade (6dB/octave), if continued over a wide frequency band, say two decades or more, will cause the excess phase lag to reach nearly 90°. A sustained attenuation rate of 40dB/decade will give nearly 180° lag, and will bring the Nyquist diagram in almost horizontally from the right, so that, with sufficient loop gain, it will pass very close to the point (1, 0). This will give a large peak in the frequency response and a very ringy step response.

In the above context it is usual to refer to the stability margins of an amplifier – the gain margin and the phase margin, as shown in Fig. 7. The gain margin is a measure of how much more feedback could be applied without oscillation occurring, and the phase margin shows how much extra internal phase lag would be necessary, at the frequency of unity loop gain, to reach the oscillation point.

The magnitudes of the stability margins that should be left in a practical amplifier design are dependent on a number of considerations, as follows:

(a) The margins, as designed, should be "comfortable", to ensure that likely production variations do not lead to trouble.

(b) The margins should satisfy (a) under all conditions of amplifier loading which it is considered necessary to take into account.

(c) Adequate margins, as in (a) and (b), should be maintained at all signal levels, not just under small-signal conditions.

(d) In many television, radar and c.r.o. amplifier designs, a step response with little or no overshoot is vital, but this is not normally the case with audio amplifiers – except, perhaps, to satisfy the demands of some reviewers and their public! Although a highly ringy step response justifiably arouses one's general suspicions about an audio amplifier design, a high-Q ring at 150kHz, say, will nevertheless not, in itself, have the slightest deleterious effect on the sound reproduction.

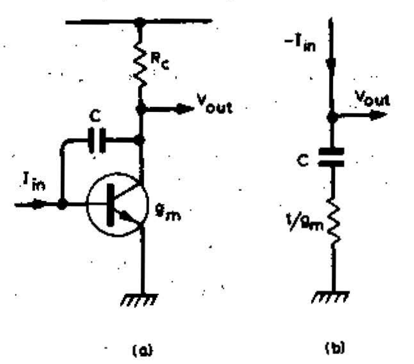


Fig. 6. *(a) Circuit not having minimum-phase shift (m.p.s.) characteristics. (b) Circuit having same amplitude response as (a), but giving less phase lag at high frequencies.*

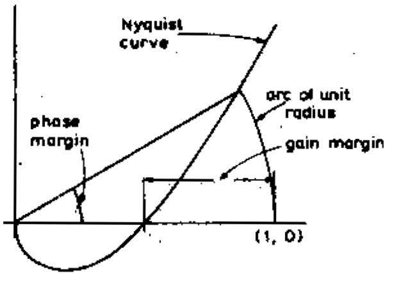


Fig. 7. *Diagram illustrating the definition of gain margin and phase margin.*

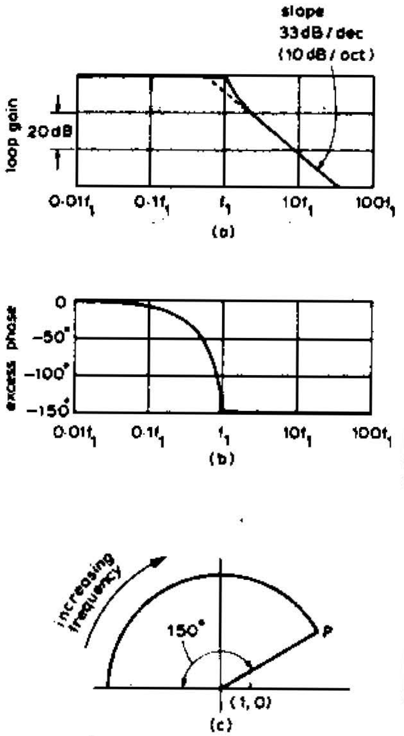


Fig. 8. *Loop-gain, phase and Nyquist diagrams for amplifier having "Bode ideal" loop-gain-attenuation characteristics.*

The argument against designing for excessive stability margins is that any audio amplifier so designed could easily be made to have less non-linearity distortion by altering the design to have more high-frequency feedback and smaller stability margins. However, with modern fast, silicon planar transistors, it is possible to obtain a superb performance, with regard to high-frequency distortion, even when the stability margins are made quite large, giving impressive-looking square-wave response. The present-day tendency is to do this.

A long time ago Bode[3] suggested that for purposes such as low-distortion audio amplifiers, a 30° phase margin is a sensible choice, and argued that a good philosophy, when practicable, is to hold the loop gain at its full value up to some chosen high frequency f_1, such as 10kHz, and then attenuate it as rapidly as possible, consistent with not exceeding 150° excess phase lag. In this way, with good design in other respects, the loop gain may be reduced to well below unity before unpredictable phase lags, due to the complex behaviour of transformers etc., become significant. He also showed that the ideal loop gain attenuation law to achieve a constant 150° phase lag above f_1 is as shown in Fig. 8(a).

If the steep rate of gain attenuation above f_1 were maintained for too long before reaching the 33dB/decade asymptotic rate, the corresponding phase curve would dip below −150°, but recover to −150° asymptotically at very high frequencies. In addition, this effect, if pronounced, would result in an amplifier having conditional stability if the loop gain were set high enough. A comparable state of affairs is shown in Figs. 4(c) and 4(d).

A characteristic of an amplifier with "Bode ideal" loop gain attenuation, assuming a flat β-network, is that the magnitude of the high-frequency response peak (6dB), and the shape of the step response, are independent of the loop-gain setting over a wide range** due to the constant 30° phase margin. This is only true if the loop gain is set high enough for the point (1, 0) to be well to the left of the point P in Fig. 8(c). I have recently set up an experimental amplifier circuit in which the loop gain is attenuated accurately at 33dB/decade, with a minimum-phase-shift network, over a frequency range of 1000:1. With the loop gain set to make the high-frequency peak occur somewhere in the central region of this frequency range, the peak is indeed of 6dB magnitude, as would be predicted from the Nyquist diagram geometry – see Figs. 8(c) and 2(a). It is also interesting to note that the step response, whose shape is almost perfectly independent of loop-gain setting over a wide range, is nearly indistinguishable from the Q=2 waveform of Fig. 2(f). There is no absolutely fixed theoretical relationship between the phase margin and the shape of the closed-loop step response, even when β is constant. Nevertheless, with reasonably "tame" Nyquist diagrams, as illustrated in this article, a 30° phase margin will always give an approximation to a single-mode ring with an effective Q-value in the region of 2.

** A feature perhaps particularly desirable in the valve era, since mutual conductances tended to fall off with valve ageing.

References

1. Scroggie, M. G., Phasor Diagrams (Iliffe 1966).
2. Nyquist, H., Regeneration Theory, *Bell System Tech. J.*, Jan. 1932, p.126.
3. Bode, H. W., Network Analysis and Feedback Amplifier Design. (van Nostrand 1945). (See p.162 re conditional stability; p.303 re gain/phase relationships; p.454 re Bode ideal attentuation characteristics.)
4. Berry, S. D., New Equipment for Outside Broadcasts, *The BBC Quarterly*, Vol. 7 No. 2, pp. 120-128 (Summer 1952).
5. Cherry, E. M. and Hooper, D. E., Amplifying Devices and Low-Pass Amplifier Design, p.501. (John Wiley 1968).

Corrections

Figures 2(e) and 2(f) should have appeared in the March issue as Figs. 6(b) and 6(c). Unfortunately, the dB scaling was produced incorrectly in Fig. 6(b), and Fig. 6(c) was inadvertently omitted.

It was stated on page 43 of the March issue that the closed-loop transfer function of an amplifier contains information about the stability margins. E. F. Good has pointed out that this statement is not strictly correct, for the stability margins are purely functions of Aβ, whereas the transfer function gives the value of A/(1−Aβ) only, i.e. (1/β)×Aβ/(1−Aβ). Aβ, and hence the stability margins as normally specified, are therefore deducible only if the value of β is known as well as that of the whole transfer function. □

P. J. Baxandall, B.Sc.(Eng.), F.I.E.E., F.I.E.R.E.

Peter J. Baxandall was educated at King's College School, Wimbledon, obtaining his degree in electrical power engineering at Cardiff in 1942. After wartime work as a radio instructor, he moved in 1944 to TRE (now RSRE), Malvern, working on microwave techniques for the first two years and then joining Professor F. C. Williams's team on electronic circuit research work.

In 1971 he set up as an independent electro-acoustical consultant, having had a strong hobby interest in this field for many years. He has since done much practical and theoretical work for British audio firms, this work involving the design of capacitor microphone circuits, transformers, oscillators, f.m. receivers, audio amplifiers and loudspeakers.

He is a member of the Audio Engineering Society and a fellow of the British Kinematograph, Sound and Television Society.

Audio power amplifier design — 4

More on feedback stability

by Peter J. Baxandall, B.Sc.(Eng), F.I.E.E., F.I.E.R.E.

The May article ended by discussing Bode's ideal loop-gain attenuation characteristic, which maintains full loop gain up to a certain frequency and then attenuates the gain as rapdily as possible consistently with not exceeding an excess phase shift of 150°. This ideal characteristic cannot normally be realized in practice, and the present article deals with some more realistic techniques.

THE 33dB/decade (10dB/octave) attenuation rate of Bode's ideal characteristic, which must be produced by minimum-phase-shift networks, is assumed to continue to indefinitely high frequencies. In practical multistage amplifiers this cannot be achieved, for the attenuation rate at very high frequencies is determined by unavoidable shunt capacitances and by transistor characteristics. Thus in the absence of circuit elements added for controlling the loop-gain attenuation, it will typically be as shown by curve 1 in Fig. 1. With suitable elements added within the forward path of the amplifier circuit, a close approximation to the Bode Ideal characteristic may be obtained up to a certain high frequency, but above this frequency, as shown by the full-line curve 2, the response inevitably follows curve 1. However, provided the 33dB/decade slope is continued for a sufficient number of dB, marked x, below unit loop gain, the resultant phase margin will not be very significantly reduced below 30°. Bode showed that in these practical cirumstances, the desired 30° margin can be retained, together with the advantage of starting the loop-gain attattenuation at a somewhat higher frequency, by adopting the characteristic shown in curve 3. The flat portion between P and Q delays the onset of further phase lag until the loop again is well below unity. This and related topics are discussed in much greater detail in Bode's book[1]. It should be noted that the definitions of phase and gain margins used by Bode are different from those illustrated in Fig. 7 of my May 1978 article. Bode takes the phase margin as applying at a loop gain which is below unity by the quoted gain-margin figure, usually 9dB. The definition I have given is also in widespread use[2,3,4] and seems more convenient for practical purposes.

It is very rare in the practical design of feedback amplifiers, either for audio or for other purposes, for any great effort to be made to follow accurately the Bode or other similar precepts for achieving full feedback up to the maximum possible frequency. Designs of this type tend to be complex and expensive, containing LCR networks to give the rapid drop in loop gain below the point A in Fig. 1, and the flat between points P and Q, together with staggered transitional-lag networks to give a close approximation to the 33dB/decade slope. Such designs have sometimes been used in critical Post Office repeater amplifier circuits.

The simplest stabilization technique is always to put in one dominant lag to attenuate the loop gain at 20dB/decade (6dB/octave), starting from a corner frequency which is sufficiently low to ensure that the loop gain is reduced to unity before the other lags inevitably present at high frequencies have produced too much further phase lag. In a multistage amplifier, the dominant lag is most straightforwardly introduced simply by putting a sufficiently large capacitor across the first stage collector load. This will produce a loop gain characteristic such as that represented by curve A in Fig. 2. The technique is in all respects sub-optimum, and it is important to note that the curve remains below curve C, obtained without any stabilizing elements, even at very high frequencies. This is the inevitable result of using any type of shunt stabilizing network, in the forward path of the amplifier, whose impedance becomes that of a capacitor at very high frequencies. The ultimate high-frequency asymptote position is lowered by *m* decibels, as shown in Fig. 2, where:

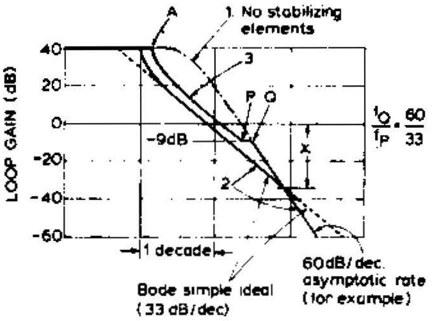


Fig. 1. *Bode simple and modified loop-gain-attenuation characteristics. See text.*

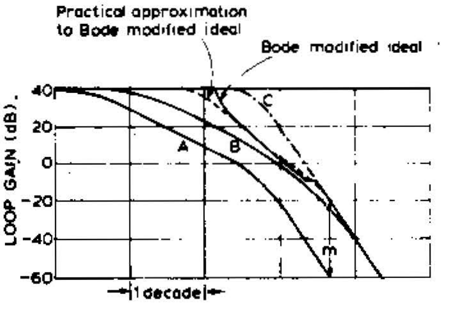


Fig. 2. *Dominant lag loop-gain-attenuation characteristics compared with Bode modified ideal. See text.*

$$m = 20 \log \frac{\textit{total shunt capacitance}}{\textit{original shunt capacitance}}$$

The introduction of any network which acts as a three-terminal potential-divider at very high frequencies will have a similar effect.

Clearly, if we wish to attenuate the loop gain in a simple 20dB/decade manner, starting at the highest possible frequency, a characteristic such as that represented by curve B in Fig. 2 must be aimed at. The simplest way to achieve this is to put a series combination of C and R across the first stage collector load. The transitional lag introduced by these elements is arranged to "flatten out" in the frequency region where the other lags come in, thus maintaining a fairly uniform rate of loop gain attenuation.

So far a single overall feedback loop has been assumed, with stabilization by means of added passive loop-gain-attenuating circuits within the forward path of the amplifier. Most modern amplifiers, however, incorporate local feedback loops as well as the main overall loop. Nyquist's criterion, in the simple form already given, is applicable to such multiple-loop amplifers only if the circuit remains stable when the overall feedback loop is broken. However, it is possible, for example, to make amplifiers in which internal positive feedback is employed to enhance the gain of part of the forward path, and such amplifiers may be unstable when the overall feedback loop is broken. An extension of Nyquist's criterion to cover such cases is described in Bode's book[1], but in many years of circuit design work involving diverse applications of feedback, I have never had to make use of this more elaborate criterion. This is because:

(a) Nearly all practical multiple-loop feedback systems employ only quite tame and stable local feedback loops.

(b) Even if the amplifier is unstable with the overall loop broken at the β-network, it is sometimes possible to break the loop at a different place, within the amplifier's forward path, leaving a stable system.

Thus, in all normal circumstances, one merely uses the closed-loop response of each internal local-feedback "sub-amplifier" as an element contributing to the total forward-path response of the complete amplifier, arranging matters so that the ordinary Nyquist stability criterion is satisfied for the overall loop.

The advantages of using local feedback, rather than purely passive networks, to tailor the internal responses of a feed-back amplifier, are often very great. One reason is that the local feedback, if applied in enlightened ways, can be exploited, like overall feedback, to reduce non-linearity distortion. Another reason is that local feedback provides a means for modifying the input and output impedances of individual stages so that they may be connected together with little interaction[5,6]. These matters will be considered in greater detail in later articles.

Stabilizing elements in the feedback arm

It has been implied so far that β is frequency-independent and that all modification of the loop-gain-attenuation characteristic to secure good stability is done in the forward path of the amplifier. There are usually advantages, however, in including one or more stabilizing elements in the feedback arm, but just what constitutes an optimum design depends upon many factors, and particularly upon the application for which the amplifier is being designed.

In feedback amplifiers for some non-audio purposes, the aim is to achieve the widest possible bandwidth of flat response, and in such cases the β-arm must have a flat response up to about the unity-loop-gain frequency. Even then it is often advantageous to add a capacitor of quite small value across the feedback resistor as shown in Fig. 3(a), sufficient to cause a little phase advance around the unity-loop-gain frequency and a reduction in the rate of attenuation of loop gain at frequencies above this. This will improve the phase margin and give a better-damped step response.

In an audio amplifier, on the other hand, if other conditions permit, there is no reason why C in Fig.3(a) should not be made much larger, giving a substantial reduction in the bandwidth of the amplifier above audio frequencies. The use of a more complex double-phase-advance network as show in Fig. 3(b) is also a possibility. In general, if the very lowest distortion up to the highest possible frequency is the requirement, the forward gain should be attenuated as little as possible and the required gradualness of loop-gain attenuation achieved as far as can be managed by arranging for the value of β to increase with rising frequency. Such designs are liable to have a very high frequency of unity loop gain, however, and it is necessary to take particular care over layout and the effects of tolerances in transistor parameters.

A feature of audio amplifiers stabilized on the Fig. 3(a) basis, with a substantial value of C, is that the step response becomes quite rounded. If, at the same time, the stability margins are rather small, the total step response is liable to be of the type shown in Fig. 4.

Very often the rapidly-increasing rate of attenuation of the forward gain at very high frequencies, with an accompanying large phase lag, prevents the possibility of carrying the technique of stabilization by manipulating the β-network very far, but from a distortion point of view it has everything in its favour.

Circles of constant gain rise

The gain of a feedback amplifier is:

$$\frac{V_{out}}{V_{in}} = \frac{A}{1-A\beta}$$

which may be written as:

$$\frac{V_{out}}{V_{in}} = (1/\beta) \times \frac{A\beta}{1-A\beta} \qquad \ldots (1)$$

Consider now the part of a Nyquist diagram shown in Fig. 5. It is clear that for points on the Nyquist curve in the region of the point (1,0), the Aβ phasor OP becomes much longer than the

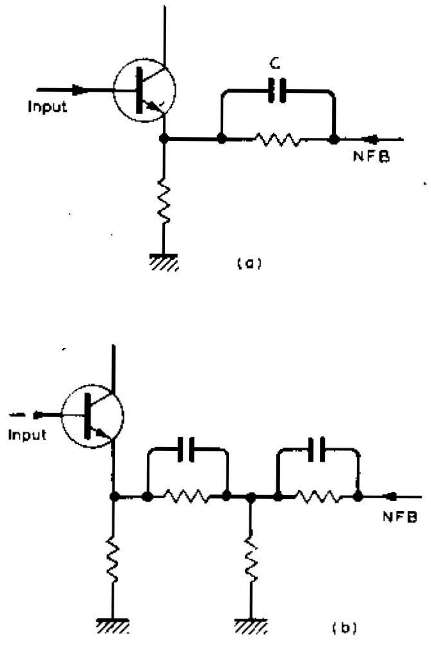


Fig. 3. *Stabilizing arrangements in the feedback arm of an amplifier. In non-audio amplifiers a low value capacitor C, in (a), can be added to improve the phase margin and give a better damped step response. In audio amplifiers C may be much larger, giving a substantial reduction in the bandwidth of the amplifier above audio frequencies. Arrangement (b) is a double-phase-advance network.*

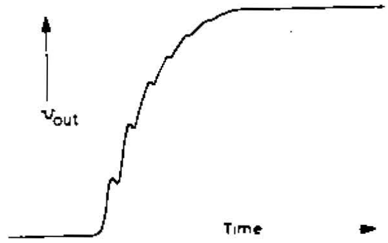


Fig 4. *Typical total step response for an audio amplifier having an arrangement as in Fig. 3 (a), with a substantial value of C, and a small stability margin.*

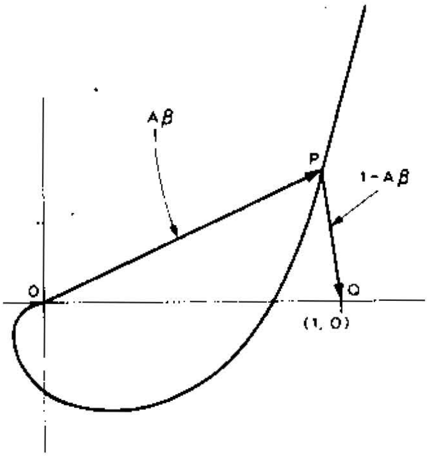


Fig. 5. *Nyquist diagram illustrating that, for points on the curve in the region of (1, 0), the phasor OP becomes much longer than the phasor PQ, so that the gain with feedback considerably exceeds 1/β and as the frequency increases the gain rises to a peak value and then falls off.*

(1—Aβ) phasor PQ, so that, from (1), the gain with feedback then considerably exceeds 1/β. Thus as the frequency increases and the point P moves down from the top of the Nyquist curve, the gain rises to a peak value and then falls off.

One may draw a set of curves on such a Nyquist diagram, each curve being for a constant ratio of OP to PQ. The simplest of these is a straight line through the point (½, 0), and if P lies anywhere on this line, OP = PQ and the gain with feedback is then exactly 1/β.

Consider now the curve for OP = 2PQ, i.e. a 6dB gain rise above 1/β. To determine this curve is a typical school geometry problem—"what is the locus of a point P such that OP = 2PQ everywhere on the curve?". The locus turns out to be a circle, centre (4/3, 0) and radius 2/3. For other ratios of OP to PQ, the locii are all circles of various radii and centre positions, as shown in Fig. 6. Note that the radial scale in this diagram is a linear one, not a decibel scale as sometimes used. This is quite satisfactory since only a small part of the complete Nyquist diagram has to be drawn.

Given the loop gain and phase information for an amplifier, the relevant part of its Nyquist diagram may be quickly sketched in on Fig. 6, and the magnitude of the high-frequency peak thereby deduced. For the Nyquist diagram shown in broken line as an example, the closed-loop response will be +3dB with respect to 1/β at f_1, will reach a maximum of +9dB at f_2, and will be −6dB at f_3, etc. If the β-network does not have a flat response at these frequencies, due allowance must be made for this in deducing the overall closed-loop response, since the diagram only gives response variations with respect to 1/β.

Mere inspection of the Fig. 6 circles gives one a pretty shrewd idea of the sort of phase margins to aim at for various types of amplifier application, bearing in mind that the step response is in practice fairly closely related to the degree of high-frequency response peaking — see Fig. 2 of the May 1978 article.

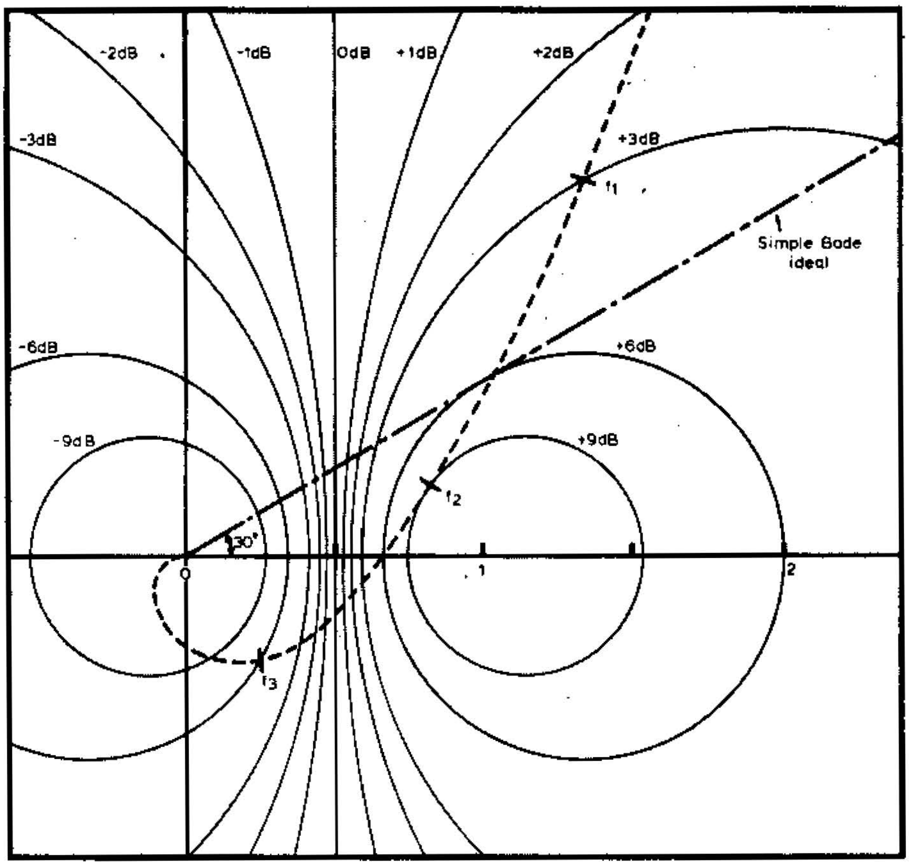


Fig. 6. *Nyquist diagram with circles of constant gain change with respect to 1/β.*

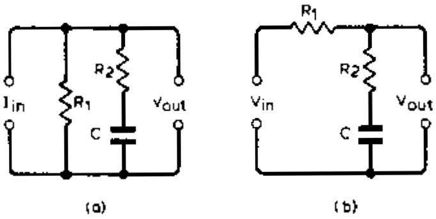


Fig.7. *Transitional lag circuits. Both circuits are equivalent and give no phase lag at zero or infinite frequencies, but contribute a phase lag which reaches a maximum value at the geometric mean of the two 'corner' frequencies. See text.*

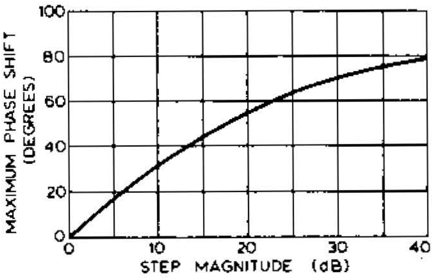


Fig. 8. *Characteristic for simple transitional lag or lead circuits.*

Maximum phase shift for transitional-lag circuit

In controlling the rate of attenuation of loop gain in feedback amplifiers, frequent use is made of transitional-lag circuits having one or other of the configurations shown in Fig. 7. The circuits are, of course, equivalent, since the combination of R_1 and I_{in} in Fig. 7(a) may be replaced by a voltage source $I_{in}R_1$ acting in series with R_1. The circuits give no phase lag at zero or infinite frequencies, but contribute a phase lag which reaches a maximum value at the geometric mean of the two corner frequencies given in the table on page 44 of the March 1978 issue. The larger the ratio of R_1 to R_2 the larger is the step in attenuation between very low and very high frequencies and the more nearly does the maximum phase lag approach 90°. Fig. 8 gives the relationship between the step magnitude in decibels and the maximum phase shift, and has been found useful for some design purposes. The graph may also be applied to the corresponding phase-lead networks sometimes used in the low-frequency stabilization of a.c.-coupled amplifiers.

Amplifier with only two lags

A particularly simple case is that of an amplifier having only two significant lags, of time-constants T_1 and T_2, in the forward path, and a frequency-independent β-network. If the low-frequency loop gain is large, then, to avoid a large high-frequency peak in the closed-loop response, T_1 and T_2 must be made very unequal, so that most of the loop-gain attenuation is done by the larger time constant without too much additional phase lag from the smaller one. A set of universal curves for this situation, calculated many years ago and recently rechecked, is given in Fig. 9. Knowing the low-frequency loop gain, the required ratio of time constants to give a specified magnitude of high-frequency peak in the response may be immediately obtained. As with Fig. 8, this data may also be applied to the corresponding low-frequency problem in an amplifier having two a.c.-

coupling time constants in the forward path.

Addition of small time constants

In many practical cases where a feedback amplifier is stabilized by the dominant lag technique, there will be one large lag plus several significant smaller lags. These smaller lags can often be satisfactorily considered as approximately equivalent to one lag of time constant equal to the sum of the individual small time constants. Some calculations relating to this are presented in graphical form in Fig. 10. It will be seen that provided the total lag introduced by the small time constants does not exceed about 40°, there is no great error in the calculated phase angle if they are taken as equivalent to a single lag of time constant equal to their sum. This procedure is very satisfactory for amplifiers having large phase margins such as 50°, and is a useful guide to approximate values, as a basis for experimentation, even when smaller phase margins are used. In this way the information of Fig. 9 may be used to some extent even when there are actually more than two time constants.

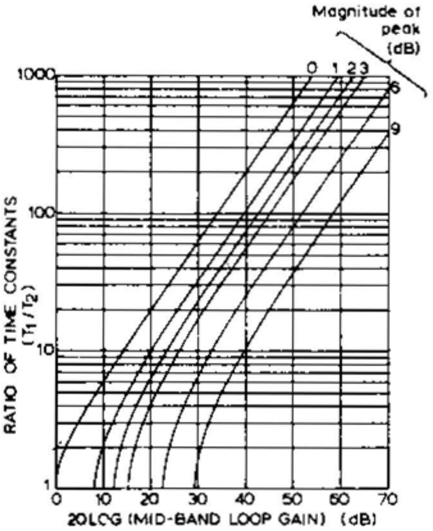


Fig.9. *Curves relating to amplifier with only two lags, or leads, in forward path, and a frequency-independent β-network.*

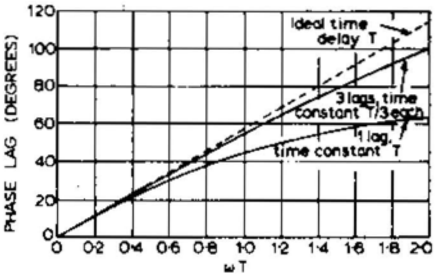


Fig. 10. *Phase lag characteristics.*

References

1. Bode, H. W., Network Analysis and Feedback Amplifier Design (van Nostrand, 1945). (See section 18.5, p.464, re the "Bode flat" in loop-gain attenuation characteristic. See section 8.8, p.157, re stability criterion for multiple-loop case.)

2. Hakim, S. S., Feedback Circuit Analysis. (Iliffe, 1966).

3. West, J. C., Textbook of Servomechanisms. (E.U.P., 1953).

4. Ghausi, M. S., Electronic Circuits. (van Nostrand Reinhold, 1971).

5. Cherry, E. M., An Engineering Approach to the Design of Transistor Feedback Amplifiers, *J. Brit. I.R.E.* Vol. 25, No. 2, Feb. 1963.

6. Cherry, E. M. and Hooper, D. E., The Design of Wide-band Transistor Feedback Amplifiers, *Proc. I.E.E.* Vol. 110, No. 2, pp.375-389, Feb. 1963. (Note, the material in refs. 5 and 6 is also covered in the book by Cherry & Hooper previously referenced.)

Audio power amplifier design — 5

Negative feedback and non-linearity distortion

Why does the low note contain the sound of the high note? — ARISTOTLE

by **Peter J. Baxandall,** B.Sc. (Eng), F.I.E.E., F.I.E.R.E.

The July article in the present series concluded the treatment of the basic techniques for achieving feedback-loop stability. Attention will now be given to the effects of negative feedback on non-linearity distortion, and it will be shown that some of the ideas involved are more subtle than is sometimes appreciated.

THE following treatment, which has gradually become clarified and extended in scope over a period of many years, will, it is hoped, enable the reader to see what the answers to questions such as the following should be:

(a) Is it a valid criticism of the use of large amounts of negative feedback that it converts moderate amounts of low-order harmonic and intermodulation distortion into a multitude of small-amplitude distortion products of high order, which may be subjectively more significant?

(b) Is it always desirable to design a feedback audio amplifier to have a nearly-level audio-frequency response before feedback is applied?

(c) Does plenty of feedback at medium audio frequencies, assuming there are no slew-rate or other overload effects, necessarily ensure that two or more signal components near the top of the audio band will give rise to negligible intermodulation products at medium frequencies?

(d) Is it important for an audio amplifier to give low distortion when signals at frequencies lying outside the audio band are fed into it?

Obviously, in many amplifier circuits, owing to the presence of capacitors or transformers, or because of insufficient bandwidth in transistors, frequency-dependent effects will have to be invoked when considering distortion mechanisms. In some practical audio circuits, however, such effects may be negligible. The following treatment will initially assume no significant frequency-dependence, and will provide a foundation of theoretical understanding which may later be extended to include the influence of frequency.

Amplifier with parabolic transfer characteristic

Consider the basic feedback amplifier configuration shown in Fig. 1. The voltage symbols represent instantaneous voltages, and each polarity marked is that which exists when the corresponding symbol has an instantaneously positive value. For the feedback to be negative, either A or β must be negative. (For a defence of the sign convention adopted, see page 41 of the March 1978 issue.) For present purposes it will be convenient to take A as being positive, so that β will be negative.

The simplest form of non-linearity to consider is that in which the transfer characteristic of the amplifying device, i.e. the graph of instantaneous output voltage (or current) against instantaneous input voltage (or current), departs from being a straight line only because of the presence of a square-law term in the corresponding equation*. Thus, referring to Fig. 1, let

$$v_{out} = Av' + \alpha(Av')^2 \qquad (1)$$

The graph of this equation is the transfer characteristic shown in Fig. 2. Plotted on this convenient basis, with equal scales on the two axes, the 45° broken line represents the slope at the origin, the actual characteristic departing from the ideal straight line by the amount $\alpha\ (Av')^2$ as shown. Because equation (1) is a quadratic equation, representing a parabola (of which only part is drawn in Fig. 2), the graph is sometimes called a quadratic transfer characteristic.

If there is no feedback in the Fig. 1 circuit ($\beta = 0$), v' becomes equal to v_{in}, and the complete circuit then has a transfer characteristic equation as in (1) but with v_{in} written for v'. Suppose now that v_{in} is a sine-wave signal given by

$$v_{in} = \hat{V}_{in}\sin\omega t \qquad (2)$$

Substituting this for v' in equation (1) gives

$$v_{out} = A\hat{V}_{in}\sin\omega t + \alpha A^2\hat{V}_{in}^2\sin^2\omega t \qquad (3)$$

(No feedback)

The first term represents the wanted fundamental output, the other term representing the second-harmonic distortion because

$$\sin^2\omega t = ½ - ½\cos2\omega t \qquad (4)$$

This elementary trigonometry formula may be illustrated graphically as in Fig. 3. (I trust that those readers highly familiar with such elementary ideas will bear with me until more interesting topics are reached — I have assumed that some readers will welcome a rather slow and basic approach.)

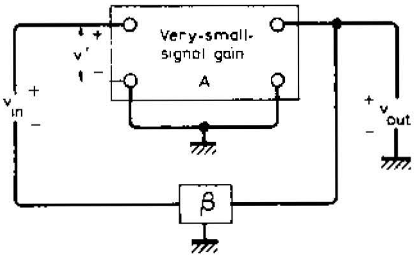


Fig. 1 *Basic feedback-amplifier configuration.*

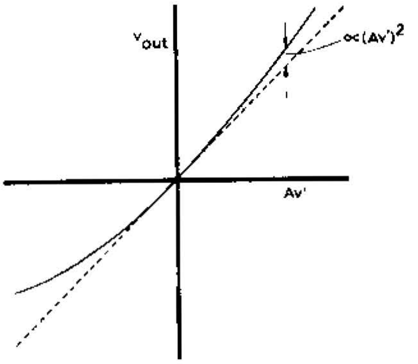


Fig. 2 *Simple parabolic, or quadratic, transfer characteristic. α is a constant determining the degree of non-linearity, and A and V′ are as in Fig. 1.*

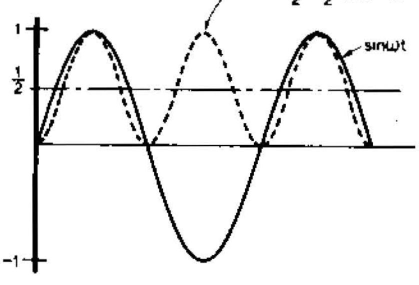


Fig. 3 *Waveforms illustrating a basic trigonometry formula.*

* It is tempting to call this equation the 'transfer function', but this usage is better avoided because the term has an almost universally accepted meaning in a somewhat different context, as explained in the March 1978 article. It is thus better to refer simply to 'the equation of the transfer characteristic'.

If (4) is substituted for $\sin^2\omega t$ in (3), it will be seen that the magnitude of the second-harmonic output voltage component is given by

$$\hat{V}_{2nd} = \tfrac{1}{2}\alpha A^2 \hat{V}_{in}^{\,2} \qquad (5)$$

(No feedback)

The magnitude of the fundamental output is given by

$$\hat{V}_{fund} = A\hat{V}_{in} \qquad (6)$$

(No feedback)

Dividing (5) by (6) and multiplying by 100 gives the percentage second-harmonic distortion as

$$\%2\text{nd} = \tfrac{1}{2}\alpha A\hat{V}_{in} \times 100 \qquad (7)$$

(No feedback)

Thus from (5) and (7), the absolute magnitude of the second-harmonic output voltage is proportional to the square of the input (or fundamental output) voltage, whereas the percentage second-harmonic distortion is linearly proportional to the input voltage itself. This is a property of any circuit or device in which square-law distortion is dominant. (It may here be mentioned that a statement such as "the distortion is proportional to the square of the output voltage" is really quite ambiguous, for "the distortion" can be taken to mean either "the distortion voltage" or "the percentage distortion". This ambiguity often appears in the literature and sometimes causes very real confusion. A plea is therefore made to authors to say what they mean!)

The problem now to be considered is the effect on distortion of making β finite in Fig. 1, i.e. applying negative feedback, still assuming a parabolic transfer characteristic for the basic amplifier. This problem may be approached from several different angles, and, as is often the case, adopting more than one viewpoint is helpful in providing a more complete understanding of the principles involved.

First of all it is possible to construct, point by point, a graph of v_{out} against v_{in} with feedback operative, and to show that it is much more nearly linear than without feedback. To do this, a particular value of v' is taken, and from equation (1), assuming A (the gain for very small signals) is known, v_{out} is calculated. Then βv_{out} is determined. Finally, with due care over signs, v_{in} is obtained from the relationship

$$v' = \beta v_{out} + v_{in} \qquad (8)$$

A little thought will show that as the magnitude of A or β is increased, the resultant transfer characteristic becomes more and more nearly a perfect straight line. With very large A or β, v_{in} becomes enormously greater than v', and the overall gain is then given very nearly by

$$v_{out}/v_{in} = -1/\beta \qquad (9)$$

(Infinite feedback)

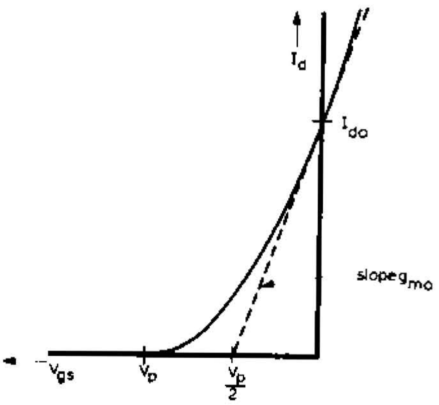


Fig. 4 *Ideal parabolic characteristic for f.e.t.*

The change from a parabolic transfer characteristic to a straight line as the loop gain is increased from zero to infinity is a smooth and gradual process. All the intermediate transfer characteristics are absolutely smooth curves, quite free from any suggestion of kinks or other blemishes. But is each one still a parabola, of lesser curvature?

The answer to the above very important question is "no", and an indication that this must be so can be obtained without actually working out the equation of the new transfer characteristic. Start with $\beta = 0$ (no feedback). With a sine-wave input at frequency f, the output will contain components at f and $2f$. As soon as β is made finite, some of the $2f$ component will be let through into the input circuit, so that the basic amplifier will now be receiving inputs at f and $2f$.

Now any device with a parabolic, or quadratic, transfer characteristic, when fed with two inputs at different frequencies, generates intermodulation products at the sum and difference frequencies – and the sum frequency in the present case is $3f$. (This arises from the fact that $(\sin\omega_1 t + \sin\omega_2 t)^2$ gives a term $2\sin\omega_1 t\sin\omega_2 t$ which is equal to $\cos(\omega_1 - \omega_2)t - \cos(\omega_1 + \omega_2)t$.) Thus, while the amplifier without negative feedback gives nothing but second-harmonic distortion on a single sine-wave input, as soon as a little feedback is applied, a third-harmonic output appears. This is not the end of the story, however, for this third harmonic, like the second harmonic, gets fed via the β-network into the input circuit, where sum and difference products are again generated. This time the sum products are at $f+3f$, which gives a fourth harmonic, and $2f+3f$, which gives a fifth harmonic. Clearly there is theoretically no end to this process – every new harmonic considered, when fed back, gives rise to harmonics of yet higher order. Before too hastily condemning feedback, however, it is wise to consider the magnitude of these effects, and also to question whether assuming a purely parabolic transfer characteristic is sufficiently closely related to the behaviour of practical devices to be of much value. Maybe they already produce comparable amounts of high-order harmonics before feedback is applied? It is evident that a fully satisfactory understanding of the problem can best be reached by a combination of theory and experiment.

Before presenting experimental results for comparison, the theory of feedback over an ideal parabolic device will be pursued further, to obtain the actual magnitudes of the various harmonics generated. The full analysis is somewhat tedious, but an outline of the approach adopted is as follows. The aim is to obtain an expression for the closed-loop transfer characteristic in the form of a power series

$$v_{out} = a_1 v_{in} + a_2 v_{in}^{\,2} + a_3 v_{in}^{\,3} + \ldots \qquad (10)$$

Then $v_{in} = \hat{V}_{in}\sin\omega t$ is put in this and the resultant harmonic magnitudes are obtained. To obtain the power series, the starting point is equation (8), the value of v' there given being substituted in equation (1). This produces a quadratic equation relating v_{in} and v_{out} which can be solved to give v_{out} as a direct function of v_{in}. The function, however, contains a square-root sign and is not in itself a power series. The binomial theorem is then used to obtain the wanted power series. Substituting $v_{in} = \hat{V}_{in}\sin\omega t$ in this series gives terms in $\sin\omega t$, $\sin^2\omega t$, $\sin^3\omega t$ etc. As illustrated in Fig. 3, $\sin^2\omega t$ produces second harmonic, and elementary extension of this principle shows that the $\sin^3\omega t$ term produces third harmonic,† and so on. The various harmonic amplitudes are thus obtained as functions of the peak input voltage, $\hat{V}_{in}$. More conveniently, however, for practical purposes, the harmonic magnitudes are expressed as functions of $\hat{V}_{out}$, on a percentage basis. This is preferable, because in assessing the performance of a feedback amplifier, one is interested in the percentages of the various harmonics present at known *output* levels. and how these vary with the amount of negative feedback used. The results of the analysis are given in Table 1. α is the "distortion constant" of equation (1), A is the amplifier forward gain for very small signal levels, and β is the feedback factor.

It is instructive to plot curves from the Table 1 formulae and to see how they compare with curves based on measurements using an approximately

† Some third-harmonic is also produced by the $\sin^5\omega t$ term, but in view of the much smaller magnitude of this contribution except at signal levels approaching the overload point, it may reasonably be neglected. The output level used in the tests is just low enough to avoid serious errors from this cause.

Table 1. Theoretic distortion formulae for feedback amplifier with parabolic forward transfer characteristic.

Harmonic number	Percentage of fundamental	Ratio of harmonic amplitudes	
		Harmonics	Ratio
2	$\frac{50\alpha\hat{V}_{out}}{1-A\beta}$	2nd : 3rd	$1 \times \frac{1-A\beta}{\alpha\hat{V}_{out}\lvert A\beta\rvert}$
3	$\frac{50\lvert A\beta\rvert\alpha^2\hat{V}_{out}^2}{(1-A\beta)^2}$	3rd : 4th	0.800 " "
4	$\frac{62.50A^2\beta^2\alpha^3\hat{V}_{out}^3}{(1-A\beta)^3}$		
5	$\frac{87.50\lvert A^3\beta^3\rvert\alpha^4\hat{V}_{out}^4}{(1-A\beta)^4}$	4th : 5th	0.714 " "
6	$\frac{131.25A^4\beta^4\alpha^5\hat{V}_{out}^5}{(1-A\beta)^5}$	5th : 6th	0.667 " "

quadratic device such as an f.e.t. Now it will be noticed that the product $\alpha\hat{V}_{out}$, raised to various powers, occurs throughout the formulae, and a value for this must be decided upon before a set of curves, such as those shown in Fig. 7, can be drawn. A convenient procedure is to choose the value of α so that the theoretical percentage second-harmonic distortion without feedback, given by the Table 1 formula as $50\alpha\hat{V}_{out}$, is the same as the measured second-harmonic distortion at the value of $\hat{V}_{out}$ adopted. This effectively matches the value of α to that of the practical circuit, and is more convenient than determining α by other means.

F.e.t. characteristics

Most text books give the following equation for the drain current, I_d, of an f.e.t. whose drain-to-source voltage, V_{ds}, is well in excess of pinch-off voltage, V_p.

$$I_d = I_{do}\left[\frac{V_{gs}}{V_p} - 1\right]^2 \qquad (11)$$

This is a parabolic relationship, as illustrated in Fig. 4, and from the geometry of this diagram it follows that

$$g_{mo} = \frac{2I_{do}}{V_p} \qquad (12)$$

An f.e.t. would therefore appear to be the ideal parabolic device for checking the distortion theory evolved above.

However, several years ago, it struck me that there would be something rather queer about a device if it accurately followed a law as given by equation (11), the reasoning being as follows. Differentiating (11) gives

$$g_m = \frac{dI_d}{dV_{gs}} = \frac{2I_{do}}{V_p} \times \left[\frac{V_{gs}}{V_p} - 1\right] \qquad (13)$$

But from (11)

$$\frac{V_{gs}}{V_p} - 1 = \sqrt{\frac{I_d}{I_{do}}}$$

and substituting this in (13) leads to

$$g_m = \frac{2I_{do}}{V_p}\sqrt{\frac{I_d}{I_{do}}}$$

Finally, using the relationship (12), this becomes

$$g_m = g_{mo}\sqrt{\frac{I_d}{I_{do}}} \qquad (14)$$

According to this equation, as the working drain current I_d is reduced, g_m falls off in proportion to the square root of I_d. Now for a junction transistor g_m varies with collector current I_c according to the relationship

$$g_m = I_c \times \frac{q}{kT} \qquad (15)$$

where q = charge of an electron, k = Boltzmann's constant, and T = absolute temperature.

Here g_m is proportional not to the square root of the collector current, but to the collector current itself, and with silicon planar transistors the relationship holds accurately in practice down to currents of less than a nanoamp. Thus, while an f.e.t. will normally have a lower g_m than a junction transistor at, say, 1mA, the more gradual fall-off in g_m with working current for the f.e.t. would, if continued, give it a much larger g_m than a junction transistor when operated at a low enough current. In view of the very basic quantities involved in equation (15), I felt this result was probably too good to be true! A measurement of g_m for an f.e.t. over a wide range of drain current was therefore made, and gave the result shown in Fig. 5. Thus it seems that a law of nature does indeed come into action to prevent the g_m of an f.e.t. exceeding that of a junction transistor. It will be seen that the steeper broken-line asymptote is fairly closely that expected for a junction transistor, and would, if continued to the right, give a g_m of nearly 40mA/V at 1mA.

Because of the above discrepancy between the usual text-book equation (11) and what is found to happen in practice, if for no other reason, one would not expect the transfer chracteristic, corresponding to Fig. 4, for an actual f.e.t., to be quite precisely parabolic. Consequently, even without negative feedback, harmonic components in addition to second harmonic must be expected to appear to some extent.

However, despite the above, the assumption that the transfer characteristic for an f.e.t. is as given by equation (11) is quite near enough to the mark to permit the magnitude of the second-harmonic distortion without feedback to be fairly accurately calculated – provided the working current is not excessively small (see Fig. 5). It may be deduced from equation (11) that

$$\%2\text{nd} = 12.5\frac{\hat{I}}{I_{dc}} \qquad (16)$$

(f.e.t. without feedback)

where $\hat{I}$ is the peak fundamental drain-current excursion and I_{dc} is the working d.c. drain current.

Equation (16) may be compared with the result for an ideal voltage-driven junction transistor, which is

$$\%2\text{nd} = 25\frac{\hat{I}}{I_{dc}} \qquad (17)$$

(Junction transistor without feedback)

In this latter case an alternative formula [1,2] is

$$\%2\text{nd} = \hat{V}_{in} \qquad (18)$$

where $\hat{V}_{in}$ is the peak signal input vol-

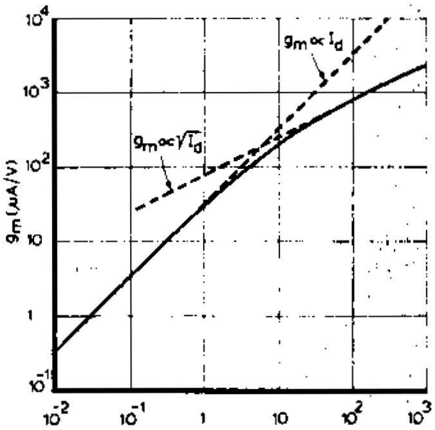


Fig. 5 *Measured mutual-conductance characteristic for an f.e.t.*

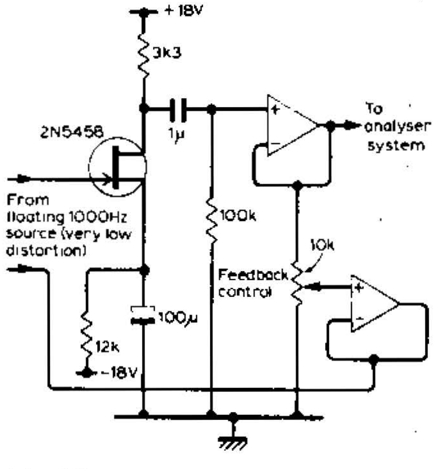


Fig. 6 *Test circuit for harmonic-distortion measurements.*

tage in millivolts. But no such delightfully simple result applies to the f.e.t.

F.e.t test circuit

The experimental circuit used for distortion measurements on an f.e.t. amplifier stage, with and without negative feedback, is shown in Fig. 6. No very expensive measuring equipment was used. The 1000Hz signal source consisted of a home-made low-distortion R-C oscillator feeding a Quad 50E amplifier, an air-cored tuned circuit purifying arrangement being connected to its floating output winding. The analyser system consisted of a parallel-T 1000Hz notch filter, whose output fed an R-C oscillator modified to function as a very highly selective amplifier feeding a c.r.o. For all measurements except second-harmonic, a passive notch circuit tuned to the second-harmonic frequency was inserted in front of the selective amplifier. Having tuned in a particular harmonic, the analyser system input was then switched to another oscillator, at the harmonic frequency, the known output of this oscillator being adjusted to give the same size of c.r.o. picture as before. With due care to avoid r.f. interference and hum problems, this set-up was both highly sensitive and of satisfactory accuracy. A test was done in which the signal source, at an enhanced level, was fed via a 3.3kΩ resistor straight to the integrated-circuit follower. The harmonic readings at the output of either integrated circuit, as the same fundamental voltage as before, were then negligible compared with those obtained with the f.e.t. in operation.

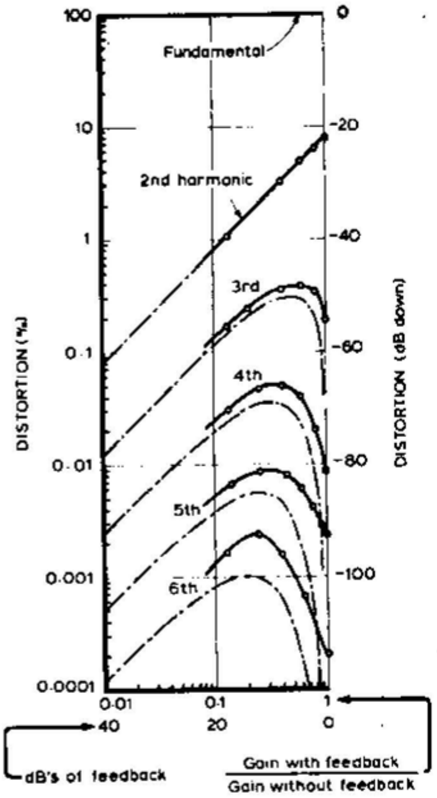


Fig. 7 *The full-line curves represent distortion measurements using the Fig. 6 set-up. The chain-dotted curves relate to calculated distortion, assuming an ideal parabolic f.e.t. characteristic as shown in Fig. 4. All curves are for a fundamental output level of 3V peak.*

Consideration of results

Fig. 7 shows, in full-line, the results of measurements using the Fig. 6 circuit, the chain-dotted curves being calculated from the formulae in Table 1. All curves relate to a fundamental output voltage of 3 volts peak. (A convenient fact is that, even with a large second-harmonic present, the peak value of the fundamental is accurately equal to half the peak-to-peak value of the total output waveform.)

The mean drain current in Fig. 6 is 1.55mA. The a.c. drain load is 3.2kΩ, giving a peak fundamental drain current, at 3 volts peak, of 0.94mA. Equation (16) above thus predicts a percentage second-harmonic distortion without feedback of $12.5\times(0.94/1.55) = 7.6\%$. It will be seen that the measured value is very close to this. As expected, however, the f.e.t. without feedback shows itself to be by no means ideally parabolic in transfer characteristic, so that appreciable amounts of higher-order harmonics are measured – though the largest of these, the third harmonic, is only 0.19% despite the quite large output level.

When feedback is applied, the magnitude of the measured third harmonic, conveniently expressed as a percentage of the constant fundamental output, at first rises, as more and more second harmonic is fed back into the input circuit to intermodulate with the fundamental voltage existing between gate and source and thus generate a sum component at third-harmonic frequency. As the feedback is further increased, the resulting improved linearity of the amplifier soon becomes the dominating influence and, when the amount of feedback is large, the third-harmonic output (at constant fundamental output) becomes directly proportional to $1/(1-A\beta)$. Similar effects occur also for the other harmonics, and it will be seen that the measured distortions, when the feedback is large, approximate closely to those calculated assuming a purely parabolic transfer characteristic. Thus, for an f.e.t. at least (though actually it applies also for a junction transistor), the main distortion mechanism for the production of third and higher harmonics, once plenty of feedback is applied, is the intermodulation one mentioned, rather than the existence of cubic and higher terms in the power series representing the transfer characteristic.

Conclusions

Some important conclusions that may be drawn from the above are:

- Even f.e.ts, used without feedback, generate high-order harmonics — and therefore, on programme, high-order intermodulation products.

- A small amount of negative feedback (e.g. 6dB) in a single-ended stage, though reducing the second-harmonic distortion, and also the total (unweighted) distortion, by about 6dB, will increase the higher-order distortion, and the quality of reproduction may well become worse as judged subjectively.

- If enough negative feedback is applied, all significant harmonics (and corresponding intermodulation products) can be reduced to a far lower level than without feedback, though the amount of feedback required to achieve this becomes larger the higher the order of the harmonic considered. (For example, referring again to Fig. 7, 16.5dB of feedback is sufficient to reduce the third harmonic to the same level as it has without feedback, whereas about 35dB is required for reducing the sixth harmonic to its no-feedback level.)

- The magnitude of harmonics of extremely high order will be increased by the application of negative feedback, no matter what practical amount of feedback is employed, but this is of no consequence if, when thus increased, they are, say, 120dB below the fundamental.

- Fig. 7, as already stated, applies at a particular output level of 3V peak in the Fig. 6 circuit, the peak drain current excursion being about 60% of the working drain current — in other words, it is high-level class A operation. When the signal level is reduced, the various harmonics fall off at different rates, as may be seen from Table 1. The percentage second-harmonic is proportional to $\hat{V}_{out}$, whereas the percentage fifth-harmonic, for example, is proportional to $\hat{V}_{out}^4$. On a logarithmic plot, as in Fig. 7, the effect of reducing the output signal level is that all the curves remain of the same shape, but each curve shifts downwards by a distance proportional to $(n-1)$, where n is the order of the harmonic, so that the spacing between the curves becomes wider. Thus at a reduced output level the higher-order harmonics rapidly become negligible.

(To be continued)

References

1. Baxandall, P. J., "Low-distortion amplifiers – Part 2", *J. British Sound Recording Association*, Nov. 1961, pp. 246-256.

2. Taylor, E. F., "Distortion in low-noise amplifiers", *Wireless World*, Aug. 1977, pp. 28-32.

Audio power amplifier design — 6

More on negative feedback and non-linearity distortion — a continuation of part 5

by **Peter J. Baxandall,** B.Sc.(Eng.), F.I.E.E., F.I.E.R.E.

Part 5 (December issue) discussed the theory of non-linearity distortion in an ideal feedback amplifier having a parabolic forward transfer characteristic. Attention is now turned to distortion in circuits using ordinary junction transistors*, having exponential transfer characteristics. The concept of "inverse distortion" is introduced, leading to a useful distortion theorem.

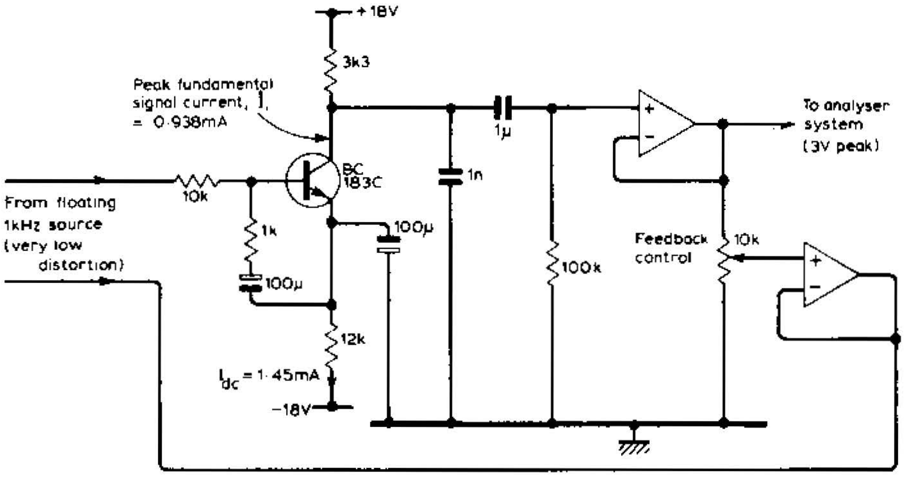

+18V
3k3
Peak fundamental signal current, $\hat{I}_c$ = 0·938mA
BC 183C
10k
From floating 1kHz source (very low distortion)
1k
100μ
100μ
12k
I_{dc} = 1·45mA
−18V
1μ
1n
100k
Feedback control
10k
To analyser system (3V peak)

Fig. 1. *Circuit used for distortion measurements.*

THE CIRCUIT USED for obtaining the experimental results presented below is shown in Fig. 1 and is the same as for the f.e.t. tests in Part 5, except for two small modifications. The 1nF capacitor was found necessary to prevent high-frequency oscillation when full feedback was applied, and the resistive attenuator in the base circuit was added to reduce loop gain, for convenience, to a similar range of values to that applying to the f.e.t. version of the circuit. The measured current gain (β_{dc} or h_{FE} of the transistor used was 580 at, an I_b of 1μA.

Throughout the measurements the fundamental output voltage was kept constant at three volts peak, corresponding to a ratio of peak signal current to direct working current of 0.647 – the same conditions as for the f.e.t. tests in Part 5. The results are shown by the full-line curves in Fig. 2, and exhibit some fascinating features when compared with the earlier f.e.t. results. A great deal of thought, both of a formally analytical and also of a more intuitive type, has been devoted to trying to understand these features, and considerable enlightenment has resulted.

A junction transistor has the great virtue, at sufficiently low values of collector current, that it follows in practice, with high accuracy, the relationship

$$I_c = I_0 \exp\frac{qV_{be}}{kT} \qquad (1)$$

where I_c is collector current, V_{be} base-to-emitter voltage, and the other symbols are constants. (The tendency always to regard junction transistors as current-operated devices, and the current gain as the basic parameter for design purposes, should be most strongly discouraged, in my opinion.)

A practical junction transistor would be expected to follow the above law much more closely than an f.e.t. would be expected to follow a parabolic law, so that there seemed good reason for thinking that the curious wiggles in the Fig. 2 curves might be theoretically explicable on the basis of equation 1.

Determining transfer characteristic

For analysis the circuit may be simplified to that shown in Fig. 3, in which the transistor is assumed to follow equation 1. It may be shown that the incremental signal input and output voltages of the circuit are related by

$$v_{out} = -R_L I_{dc}\left[\exp\frac{qv_{in}}{kT}\times\exp\frac{q\beta v_{out}}{kT}-1\right] \qquad (2)$$

where q is the electronic charge (1.60×10^{-19} coulomb) k Boltzmann's constant (1.38×10^{-23} joules/deg C) and T absolute temperature. To be able to calculate the harmonics in v_{out} when v_{in} in equation 2 is put equal to $\hat{V}_{in}\sin\omega t$, the relation must be expressed in the form of a power series:

$$v_{out} = a_1 v_{in} + a_2 v_{in}^2 + a_3 v_{in}^3 + a_4 v_{in}^4 + \ldots \qquad (3)$$

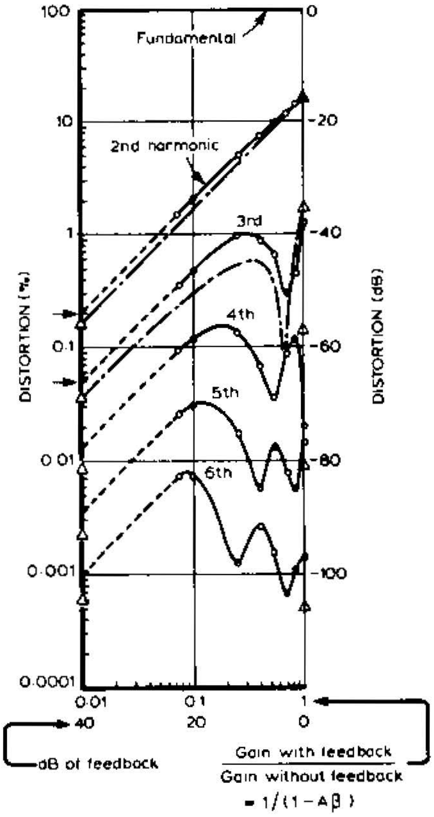


Fig. 2. *Measured and calculated results for the Fig. 1 circuit.*

*Sometimes called bipolar junction transistors or b.j.t. because their operation involves both polarities of charge carrier. The usual type of f.e.t. is also a junction device, but it is unipolar because only one polarity of charge carrier is involved.

The values of the coefficients a_1, a_2, a_3 etc may be found by using Maclaurin's theorem, which says

$$a_1 = \left[\frac{dv_{out}}{dv_{in}}\right]_{vin=0}$$

$$a_2 = \frac{1}{2!}\left[\frac{d^2v_{out}}{dv_{in}^2}\right]_{vin=0}$$

$$a_3 = \frac{1}{3!}\left[\frac{d^3v_{out}}{dv_{in}^3}\right]_{vin=0}$$

By successively differentiating equation 2 and putting $v_{in}=0$ in the resultant expressions, the coefficients may thus be determined. Unfortunately the algebra rapidly becomes cumbersome, and being no mathematician, I gave up after determining the first three coefficients, which are

$$a_1 = \frac{A}{1-A\beta} \qquad (4)$$

$$a_2 = \frac{1}{2!}\frac{q}{kT}\frac{A}{(1-A\beta)^3} \qquad (5)$$

$$a_3 = \frac{1}{3!}\left(\frac{q}{kT}\right)^2 A\left[\frac{1}{1-A\beta} - \frac{3|A\beta|}{(1-A\beta)^2} + \frac{3|A\beta|^2}{(1-A\beta)^3} - \frac{|A\beta|^3+3|A\beta|}{(1-A\beta)^4} + \frac{3|A\beta|^2}{(1-A\beta)^5}\right| \qquad (6)$$

In these equations β is positive and $A=-g_mR_L$, where g_m is the transistor mutual conductance when $v_{in}=v_{out}=0$ and the collector current is I_{dc}.

Determining the harmonics

Knowing the value of v_{in} ($=\hat{V}_{in}\sin\omega t$), as a function of the amount of feedback in use, for the output level of 3V peak adopted, the output harmonic magnitudes may be calculated from equation 3 on the assumption that only the square-law term is responsible for the second harmonic and only the cubic term for the third harmonic. Because the output level is large, this simplifying assumption leads to appreciable, though not unduly gross, errors, and for better accuracy the production of some second harmonic due to the presence of a fourth-power term needs to be taken into account, etc. A fairly high output level was adopted in the experiments to make the high-order harmonics sufficiently large for straightforward measurement, i.e. well over 0.0001%.)

The calculated second and third-harmonic curves are shown chain-dotted in Fig. 2, and lie somewhat below the measured curves because of the above simplifying assumption. The reasons for other detailed differences will become apparent later on.

In view of the + and − signs in front of the terms in equation 6, and on the supposition that the expressions for a_4, a_5 etc. will contain even more terms of both signs, one can at least say that it is hardly surprising that the measured curves for the higher-order harmonics in Fig. 2 are of a more complex type.

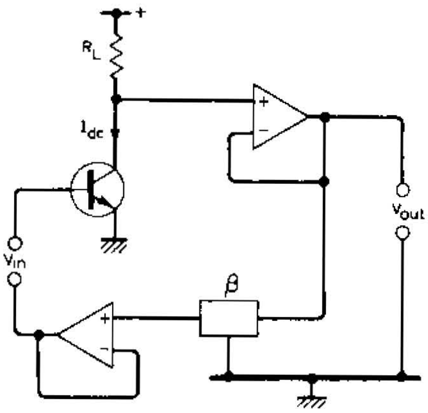


Fig. 3. *Simplified version of Fig. 1 with d.c. bias arrangements omitted.*

Alternative approach

The method of analysis presented above basically involves determining the transfer characteristic for the complete feedback amplifier, and then calculating the harmonic magnitudes when a sine-wave input is handled via this transfer characteristic. The shape of the overall transfer characteristic changes as the amount of feedback is altered, resulting in the observed variation in the magnitudes of the various harmonics. It should be emphasized that no intermodulation concept is involved in this approach when the input to the complete circuit is a single sine-wave signal.

An alternative approach, which is very helpful in providing further insight, involves thinking simply in terms of the invariant transfer characteristic of the forward amplifying device. Intermodulation effects then *do* have to be taken into account, for the forward amplifying device receives inputs from both the sine-wave input signal and also via the β-network from the amplifier output, the last-mentioned contribution containing harmonics which intermodulate with the fundamental and with each other.

In particular, the second harmonic and fundamental intermodulate to produce a component at third-harmonic frequency, and careful consideration of the waveform polarities involved shows that this third-harmonic component is in antiphase with that produced by straightforward third-harmonic distortion of the fundamental. The null in the curve for total third-harmonic distortion thus occurs when the amount of feedback is such as to make these oppositely-phased third-harmonic components of equal magnitude. The fact that in the measured third-harmonic curve, a minimum rather than a perfect null is observed, is believed to be because slight phase errors in the experimental circuit prevented the two third-harmonic components from being exactly in antiphase.

A further intermodulation effect is that the fundamental and third harmonic intermodulate to produce a second-harmonic component which, though of considerably smaller magnitude than that produced by straightforward second-harmonic distortion of the fundamental, nevertheless slightly modifies the shape of the second-harmonic curve.

The percentage of second harmonic generated within the transistor, at constant output, is proportional to the third-harmonic voltage fed back into the base circuit, but the percentage output distortion is reduced (1−Aβ) times relative to this by negative feedback. For working conditions to the left of the null in the third-harmonic curve, this intermodulation-generated second harmonic is in the same phase as that produced by straightforward second-harmonic distortion. Its magnitude at the amplifier output, with enough feedback to bring the working point onto the approximately constant-slope part of the third-harmonic curve, is such as to lift the position of the second-harmonic curve by a constant distance, and the calculated spacing is of the order shown.

It is instructive to compare the Fig. 2 curves with those of Fig. 7 in Part 5. There are two basic differences (a) the f.e.t. curves show no nulls or minima; and (b) the measured f.e.t. second-harmonic curve does not exhibit the departure from linearity evident in the Fig. 2 curve. The reason for (a) above is believed to be that, for the f.e.t. specimen used, the harmonic-distortion-generated third-harmonic component was in phase, rather than in antiphase, with the component generated by intermodulation. The high-order terms in the transfer characteristic for an f.e.t., unlike those for a junction transistor, seem to vary from one specimen to another – the one used for the Fig. 7 (Part 5) results had been selected for low third harmonic. It may well be that some other speciments would give curves with nulls, but this has not been investigated.

The reason for difference (b) above is simply that the signal level was too low to make the effect noticeable. Though the f.e.t. and the junction transistor were both worked at the same ratio of peak signal to direct working current, the f.e.t., because of its different type of transfer characteristic, gave less second-harmonic distortion in the absence of feed-back – see equations 16 and 17 in Part 5. On turning up the signal level in the f.e.t. circuit for 4V rather than 3V peak output, an appreciable departure of the second-harmonic curve from linearity was observed.

High-feedback theory

It is a characteristic of the Fig. 2 curves that all their complex features disappear when enough feedback is applied, and this fact suggests that maybe the high-feedback parts of the curves at the left could be calculated in a manner devoid of the above complications. This indeed turns out to be the case, and it is thought that appreciation of this fact is of considerable engineering value, for in the majority of practical applications one is really only interested in the performance with plenty of feedback applied.

Any amplifier, without feedback, can in principle be made to give a perfectly sinusoidal output voltage, at a specified level, by feeding an appropriately distorted waveform to its input. With negative feedback applied, this same totally undistorted output voltage can be maintained if V_{in} (Fig. 4) is arranged to contain the necessary distorted error voltage, as above, plus some extra fundamental to cancel the fundamental being injected negatively into the input circuit via the β-network. (With undistorted output, the feedback voltage is, of course, also perfectly sinusoidal.) Thus, as β is increased, V_{in} has to supply a constant-amplitude harmonic spectrum plus an increased amount of fundamental. The magnitude of the required fundamental input, for the specified constant output voltage V_{out}, is given by the usual feedback formula.

$$\frac{V_{out}}{V_{in}}=\frac{A}{1-A\beta}$$

which for the present purpose is more conveniently arranged as

$$V_{in}=V_{out}\frac{1-A\beta}{A}$$

Since the harmonic part of the input is quite constant, the percentage input distortion is inversely proportional to the amount of fundamental input voltage, i.e. it is proportional at $1/(1-A\beta)$, and this applies at every harmonic frequency. It also applies whether the amount of feedback is large or small.

It is thus seen that the distortion situation for a feedback amplifier is really very much simpler when viewed on this basis of percentage input distortion for a pure output, than when considered on the more usual basis of the output distortion for a pure sinusoidal input. At this point the reader may well object that, while it may indeed be easier to consider the feedback mechanism on this basis, the concept is artificial and not related to the way amplifiers are used in practice. The utility of the approach, however, lies in the fact that, *provided there is plenty of feedback*, the distortions become practically identical whether expressed on a distorted-input/pure-output basis, or on the usual distorted-output/pure-input basis. Thus if the percentage distortion with no feedback is calculated on a pure-output/distorted input basis – which turns out to be relatively easy – then the distortion with plenty of feedback applied, expressed in the customary manner, is equal to the just-mentioned no-feedback percentage divided by $(1-A\beta)$, the output level being kept constant. This applies both to total harmonic distortion and also to all individual harmonics of practical significance, provided only that the amount of feedback is sufficiently large. For the working conditions relevant to Fig. 2, or Fig. 7 of Part 5, it is evident that 20 to 26dB of feedback would be "sufficiently large."

It is now necessary to justify the statement that the distortion with plenty of feedback is practically the same whether expressed on a distorted-input/pure-output basis, or on a distorted-output/pure-input basis. With reference to Fig. 4, consider the state of affairs when V_{in} is of pure sine waveform, suitably adjusted in magnitude to maintain a constant output voltage no matter how much feedback there is. With no feedback, V' will be equal to V_{in} and will be sinusoidal, V_{out} being highly distorted. As the amount of feedback is increased, V_{out} becomes more and more nearly sinusoidal, which requires that the V' waveform must approximate more and more closely to that specific highly-distorted waveform, characteristic of the particular forward amplifier, which will make it deliver a perfectly sinusoidal output. The whole of the distortion in V' – call it V_{dist} – is supplied from the β-network, since V_{in} is pure. When the amount of feedback is large, the fundamental output from the β-network, injected into the input circuit, is very nearly equal in magnitude to V_{in}. Hence the percentage distortion in the output from the β-network, and therefore also in the amplifier output voltage, which feeds the β-network, is very nearly $(V_{dist}/V_{in})\times 100\%$. If now, with this large amount of feedback applied, a slight harmonic content is introduced into the V_{in} waveform so as to make the *output* perfectly sinusoidal, neither the magnitude of V_{in}, nor the harmonic content of the V' waveform, will change by more than a tiny amount, so that the distortion will still be given quite closely by $(V_{dist}/V_{in})\times 100\%$. Thus the larger the amount of feedback, the more nearly does the percentage output distortion for pure input become equal to the percentage input distortion for pure output.

Another argument to support the statement that the percentage distortions, with a large amount of feedback applied, are virtually the same when expressed on either basis mentioned above, is as follows. Referring to Fig. 4 again, suppose V_{in} contains the necessary harmonics to make V_{out} perfectly sinusoidal. Now, with these input harmonics still present, imagine that we add a further set of input harmonics, each of equal magnitude to, and in antiphase with, the corresponding harmonic already there. The result will be to cancel all the input harmonics, but introduce harmonics into V_{out}. If the harmonics thus introduced into V_{out} are simply the result of the faithful amplification of the additional set of harmonics fed in, then it follows that the percentage distortions must be the same whether considered on a distortionless input or a distortionless output basis. Whether this is nearly enough the case for practical purposes depends on how low is the intermodulation distortion introduced by the complete feedback amplifier when fed with these small-amplitude additional input harmonics in the presence of a large fundamental input, and clearly the more feedback there is, the less significant will be any "false harmonics" introduced by intermodulation – intermodulation between the fundamental and the second harmonic might introduce some third harmonic, for example.

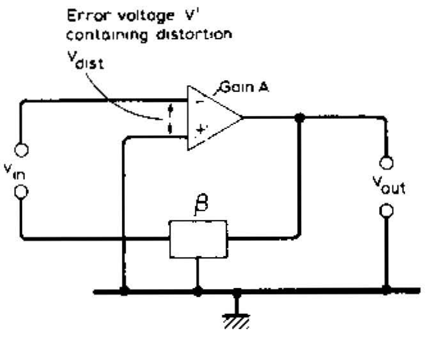


Fig. 4 *Basic feedback amplifier configuration.*

Thus, once again, the conclusion is reached that, *provided there is enough feedback*, harmonic-distortion percentages will be very nearly the same whether expressed on the normal pure-input basis, or on the inverse basis of input distortion for pure output.

A distortion theorem

The ideas discussed above may be formulated as a distortion theorem, applicable to total and to individual-harmonic distortion:

"The percentage harmonic distortion in the output of an amplifier having a large amount of feedback and a sine-wave input, is very nearly equal to the percentage input distortion for distortionless output without feedback, divided by $(1-A\beta)$."

The usefulness of this theorem is dependent on knowing the input distortion required to give distortionless output without feedback, for common amplifying devices and circuits, but fortunately the theory required is relatively simple. Such distortion can be termed "inverse distortion."

Junction transistor inverse-distortion theory

A simple single-ended junction-transistor stage will be considered first, the transistor being assumed to follow equation 1. When V_{be} is such as to cause I_c to vary sinusoidally,

$$I_{dc}+\hat{I}_c\sin\omega t=I_o\exp\frac{qV_{be}}{kT} \qquad (7)$$

in which V_{be} has the appropriate special waveform which it is desired to find. When $\hat{I}_c\sin\omega t$ passes instantaneously through zero

$$I_{dc}=I_o\exp\frac{qV_{dc}}{kT} \qquad (8)$$

where V_{dc} is the value of V_{be} required to establish the mean collector current I_{dc} in the absence of a signal input. Equation 7 may be written

$$\log_e\left[\frac{I_{dc}+\hat{I}_c\sin\omega t}{I_o}\right]=\frac{qV_{be}}{kT}$$

from which may be derived

$$V_{be}=\frac{kT}{q}\left[\log_e\frac{I_{dc}}{I_o}+\log_e(1+\frac{\hat{I}_c}{I_{dc}}\sin\omega t)\right] \qquad (9)$$

But from equation 8,

$$\log_e\frac{I_{dc}}{I_o}=\frac{qV_{dc}}{kT},$$

so that equation 9 becomes

$$V_{be}=V_{dc}+\frac{kT}{q}\log_e(1+\frac{\hat{I}_c}{I_{dc}}\sin\omega t)$$

We now use the fact that

$$\log_e(1+x)=x-x^2/2+x^3/3-x^4/4+\ldots$$

which leads to

$$V_{be}=V_{dc}+\frac{kT}{q}\left[\frac{\hat{I}}{I_{dc}}\sin\omega t-½\left(\frac{\hat{I}}{I_{dc}}\right)^2\sin^2\omega t\right.$$

$$\left.+⅓\left(\frac{\hat{I}}{I_{dc}}\right)^3\sin^3\omega t-¼\left(\frac{\hat{I}}{I_{dc}}\right)^4\sin^4\omega t+\right] \qquad (10)$$

On the assumption that $\hat{I}/I_{dc}$ is not so large that, for example, the second harmonic generated by the $\sin^4\omega t$ term is large enough to cause serious error, equation 10 yields harmonic percentages as given in the middle column of Table 1. Since $g_m=qI_{dc}/kT$ and $\hat{I}=g_m\hat{V}_{in}$, we may replace $\hat{I}/I_{dc}$ by $q\hat{V}_{in}/kT$. At 290 K, which is approximately relevant to low-level stages at least, kT/q is 25mV. These facts enable the results in the right-hand column of Table 1 to be calculated.

Table 1. **Theoretical input distortion percentages for pure sinusoidal output from ideal junction transistor without feedback.**

Harmonic number	Distortion %	Distortion (%), alternative formulae for 290K (V_{in} in mV)
2	$25(\hat{I}/I_{dc})$	$\hat{V}_{in}$
3	$8.33(\hat{I}/I_{dc})^2$	$1.33\times10^{-2}\hat{V}_{in}^2$
4	$3.13(\hat{I}/I_{dc})^3$	$2.00\times10^{-4}\hat{V}_{in}^3$
5	$1.25(\hat{I}/I_{dc})^4$	$3.20\times10^{-6}\hat{V}_{in}^4$
6	$0.521(\hat{I}/I_{dc})^5$	$5.33\times10^{-8}\hat{V}_{in}^5$

Comparison with "normal" distortion

It is instructive to compare the Table 1 results with those giving the *output* distortion for an ideal sine-wave-driven junction transistor without feedback. Referring to equation 1, put $V_{be}=V_{dc}+\hat{V}_{in}\sin\omega t$, where V_{dc} is a direct bias voltage. This leads to

$$\frac{i_c}{I_{dc}}=\exp\frac{q\hat{V}_{in}\sin\omega t}{kT}-1 \qquad (11)$$

where i_c is the instantaneous signal component of collector current and I_{dc} the collector current when $\hat{V}_{in}\sin\omega t=0$. This time the required matematical fact is that

$$\exp x=1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\frac{x^4}{4!}+\ldots$$

which gives

$$\frac{i_c}{I_{dc}}=\frac{q}{kT}\hat{V}_{in}\sin\omega t+\frac{1}{2}\left(\frac{q}{kT}\right)^2\hat{V}_{in}{}^2\sin^2\omega t$$

$$+\frac{1}{6}\left(\frac{q}{kT}\right)^3\hat{V}_{in}{}^3\sin^3\omega t$$

$$+\frac{1}{24}\left(\frac{q}{kT}\right)^4\hat{V}_{in}{}^4\sin^4\omega t+\ldots \qquad (12)$$

The harmonic percentages may then be evaluated on the same basis as for Table 1, as functions of both $\hat{V}_{in}$ and $\hat{I}/I_{dc}$, since

$$\hat{V}_{in}=\frac{\hat{I}}{I_{dc}}\times\frac{kT}{q}$$

Substituting this in equation 12 gives

$$\frac{i_c}{I_{dc}}=\frac{\hat{I}}{I_{dc}}\sin\omega t+\frac{1}{2}\left(\frac{\hat{I}}{I_{dc}}\right)^2\sin^2\omega t+$$

$$\frac{1}{6}\left(\frac{\hat{I}}{I_{dc}}\right)^3\sin^3\omega t+\frac{1}{24}\left(\frac{\hat{I}}{I_{dc}}\right)^4\sin^4\omega t+\ \ldots \qquad (13)$$

From equations 12 and 13 have been calculated the results given in Table 2. As before it is assumed that $\hat{I}/I_{dc}$ is small enough to ensure that a negligible portion of the total second-harmonic generated arises from the presence of the $\sin^4\omega t$ term, etc. However, since the terms in equation 10 fall off in magnitude with increasing order less rapidly than in equation 13, a given high value of $\hat{I}/I_{dc}$ causes larger errors in the inverse-distortion figures of Table 1 than it does under the conditions of Table 2.

Table 2. **Theoretical output distortion percentages for pure sinusoidal input voltage to ideal junction transistor without feedback.**

Harmonic number	Distortion (%)	Distortion (%), alternative formulae for 290 K($\hat{V}_{in}$ in mV)
2	$25(\hat{I}/I_{dc})$	$\hat{V}_{in}$
3	$4.17(\hat{I}/I_{dc})^2$	$6.67\times10^{-3}\hat{V}_{in}^2$
4	$0.521(\hat{I}/I_{dc})^3$	$3.33\times10^{-4}\hat{V}_{in}^3$
5	$0.0521(\hat{I}/I_{dc})^4$	$1.33\times10^{-7}\hat{V}_{in}^4$
6	$0.00434(\hat{I}/I_{dc})^5$	$4.44\times10^{-10}\hat{V}_{in}^5$

Inverse distortion for parabolic device

For an amplifier having the general parabolic transfer characteristic given by equation 1 of Part 5, repeated here as equation 14, the formulae for input distortion for pure

$$v_{out}=Av'+\alpha(Av')^2 \qquad (14)$$

sinusoidal output without feedback are given in Table 3, middle column. For an ideal f.e.t. there is the restriction that the bottom of the parabola must lie on the zero-drain-current axis, as shown in Fig. 4 of Part 5, and it then follows that $\alpha\hat{V}_{out}$ may be replaced by $¼(\hat{I}/I_{dc})$, giving the formulae in the right-hand column of Table 3. (This substitution may also be made for αV_{out} in Table 1 of Part 5 when applied to an ideal f.e.t.)

Table 3. **Theoretical input distortion for pure output for general parabolic device, and f.e.t. without feedback.**

Harmonic number	Distortion (percentage)	
	General parabolic device	Ideal f.e.t.
2	$50\alpha\hat{V}_{out}$	$12.5\hat{I}/I_{dc}$
3	$50\alpha^2\hat{V}_{out}{}^2$	$3.12(\hat{I}/I_{dc})^2$
4	$62.5\alpha^3\hat{V}_{out}{}^3$	$0.977(\hat{I}/I_{dc})^3$
5	$87.5\alpha^4\hat{V}_{out}{}^4$	$0.342(\hat{I}/I_{dc})^4$
6	$131\alpha^5\hat{V}_{out}{}^5$	$0.128(\hat{I}/I_{dc})^5$

Comparing the right-hand column of Table 3 with the middle column of Table 1, the input harmonics for the f.e.t., at a given $\hat{I}/I_{dc}$ are smaller and decay more rapidly with increasing order than for a voltage-driven junction transistor. However, in many practical feedback circuits, this apparent disadvantage of the junction transistor will be more than compensated by the fact that it has a much higher mutual conductance, giving a higher feedback loop gain and thus reducing all significant harmonics to a lower level than for the f.e.t.

With regard to the f.e.t. investigation of Part 5, dividing the figures determined from the right-hand column of Table 3 by 100 gives points on the left-hand vertical axis of Fig. 7 in Part 5 which coincide with the intercepts of the chain-dotted curves.

Relationship to experimental results

The distortion theorem formulated above may be used to calculate quickly and easily, the approximate output distortion for a single junction transistor stage having, say, 40dB of feedback, for $\hat{I}/I_{dc}=0.647$ as used in the experiments with the Fig. 1 circuit. The no-feedback inverse-distortion figures are determined from the middle column of Table 1, and are divided by 100 to give the predicted distortion with feedback. The values obtained are indicated by triangles on the left-hand vertical axis of Fig. 2.

As already explained, the Table 1 formulae assume $\hat{I}/I_{dc}$ is small enough for the amount of second harmonic produced by the 4th and 6th power terms in the power series to be ignored, etc. With $\hat{I}/I_{dc}$ as high as 0.647, there is, however, an appreciable error due to this cause. Calculation shows that the amounts of inverse second harmonic arising from the $\sin^4\omega t$ and $\sin^6\omega t$ terms in equation 10 are approximately 21% and 4% of the amount due to the $\sin^2\omega t$ term, the amounts produced by even higher-order terms being relatively negligible. Thus the true second-harmonic figure would be expected to be about 26% higher than that calculated from Table 1, the error becoming rapidly smaller with reduction in signal level. This more accurate theoretical prediction is indicated by the upper arrow at the left of Fig. 2, and ties up well with the broken-line extrapolation of the measured curve.

At the zero-feedback end of the Fig. 2 curves the simple theoretical distortion values are given by the middle column of Table 2, for $\hat{I}/I_{dc}=0.647$, and the values obtained are indicated by the triangles on the right-hand vertical axis of Fig. 2. As already stated, the errors under the Table 2 conditions, caused by working at a rather high signal level, are much less than for Table 1, but there are other causes of errors to be considered. Nevertheless, the calculated second-harmonic percentage agrees quite well with the experimental value. The shape of the second-harmonic curve can thus be explained in terms of the increasing effect of the high-order terms in the power series as the amount of feedback is increased – an alternative but equally sound explanation to that previously given involving intermodulation within the forward amplifier.

The theoretical zero-feedback points, marked by triangles, for harmonics higher than the second do not agree well with the measured values. The reason for this is believed to be that when the Fig 1 circuit is set for nominally zero feedback, a small but finite amount of negative feedback is effectively still in operation, mainly because of the presence of finite resistance in the base circuit. If this resistance, including $r_{bb'}$, totals 1.2kΩ, and assuming β_{ac} or h_{fe} of 500, it is equivalent to a resistance of 2.4Ω in the emitter lead,l causing $1/(1-A\beta)$ to be effectively 0.88 when set for nominally 1.0. To allow for this, the extreme right-hand plotted points on all the experimental curves should be moved to the left to $1/(1-A\beta)=0.88$. The effect of the 2.4Ω is negligible because of the logarithmic scale used in Fig. 2, except toward the right-hand side of the curves. With the curves thus shifted to the left, it seems reasonable to suppose that continuing the patterns of undulations already established, towards the $1/(1-A\beta)=1.0$ axis, would bring the curves approximately to the theoretical values marked by triangles.

When allowance is made for the production of third harmonic by the $\sin^5\omega t$, $\sin^7\omega t$ and $\sin^9\omega t$ terms in equation 10, the magnitude of the third-harmonic distortion voltage is increased by approximately 32%, raising the calculated value to that indicated by the lower arrow in Fig. 2, which again then ties up well with the broken-line extrapolation of the measured curve. The corresponding tedious calculations have not been done for the 4th, 5th and 6th harmonics, but it seems probable that they, too, would raise the levels of the points marked by triangles to give reasonable agreement with the broken-line, 45°, extrapolations of the measured curves.

There is a further small point which must now be mentioned. In Table 2, for a junction transistor without feedback driven by a sine-wave voltage, the factor $\hat{I}/I_{dc}$ appears. $\hat{I}$ is the peak value of the fundamental current and I_{dc} is the value of the collector current at the moment when the input signal voltage goes through zero. It would also be the quiescent current, if the transistor were operated at fixed bias voltage, and the mean current with the signal applied would then be larger than I_{dc} because of the rectifying action. However, the mean current is prevented from rising significantly when the signal is present in the Fig. 1 circuit, owing to the virtually-constant current in the 12kΩ emitter resistor. This results in the distortion being higher than the simple theory predicts. The fact that the measured second-harmonic curve goes through the 16% point predicted by Table 2 at its top end is thus fortuitous. The effect just mentioned tends to raise the level of the point, whereas the fact that there is a little feedback in action, even when the control is set for nominally zero feedback, tends to lower it. Once there is plenty of feedback in action both these effects become negligible.

It can thus be concluded generally that provided plenty of feedback is assumed right at the beginning, the more awkward parts of the theory outlined in this article, though academically interesting, do not need to be taken into account for design purposes. □

(To be continued)