

Ref. 64

Series Feedback

GEORGE FLETCHER COOPER

The classical formula for feedback conceals a factor which may lead to transient intermodulation, especially in transistor amplifiers.

$$\mu_f = \mu / (1 - \mu\beta)$$

I AM UTTERLY BORED with this quotation which I have read and written countless times. It is not even completely true. These statements are calculated to attract the attention of the reader and to serve as a rather violent introduction to some consideration of the transient behavior of feedback amplifiers. Just how important the effects we shall be considering are in the actual use of amplifiers is rather difficult to assess because in general anyone who is seriously concerned with really high quality reproduction makes use of equipment which is never fully extended. This means that there is no need to examine critically the limitations of the equipment. The only thing is that this is not good engineering: the system should do what you want it to do and very little more. After all, you do not run a 10-ton truck just to bring the groceries home, although I'm sure there is at least one man in Detroit trying to do the equivalent for the 1964 models. We shall be facing the problem more seriously soon, as we try to get high fidelity in transistor amplifiers. It is going to hurt, because we moved up to decent quality with tubes by easy stages with the professional equipment leading the way: transistor amplifiers must come right up to our present standards at once if we are to take them seriously.

Transistor power amplifiers of the 10-20 watt class are not too difficult to design and build. The problem is achieving low distortion. Unless the transistors used in the output stage are extraordinarily expensive the designer is faced with a frequency cut-off somewhere in the range between 4000 and 10,000 cps. In general, the less power he asks for, the higher the limiting frequency. This limiting frequency dominates the whole design of the feedback loop and the over-all result is that we are forced to design transistor amplifiers very much nearer to the actual requirements because we must pay very much higher penalties for overdesign.

We must now consider what happens when we apply a large signal to a feedback amplifier. Since we are engineers we shall use a square-wave input to begin with but the implications of other waveforms will be examined later. We shall first of all disconnect the feedback so that we get from the input shown in (A) of Fig. 1 the output shown in (B).

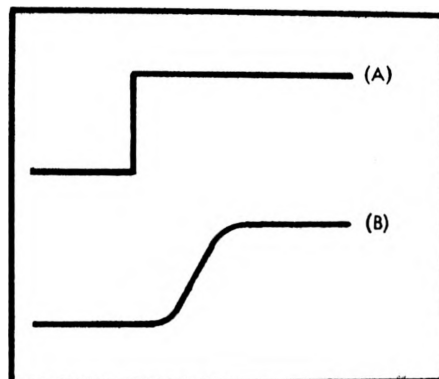


Fig. 1. The input square wave (A) is delayed and rounded off (B) before it reaches the output.

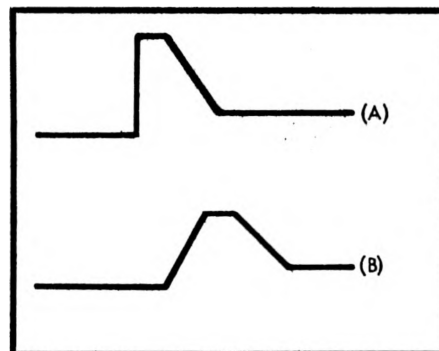


Fig. 2. As the initial signal goes through the amplifier and is fed back, the first passage results in an input (A) with a corresponding output (B).

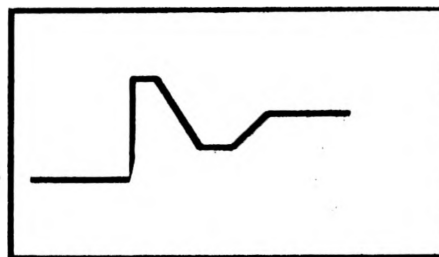


Fig. 3. When the signal shown in (B) of Fig. 2 gets through the feedback path (second time around) the total input has the form shown.

Because the amplifier has a limited bandwidth the output lags a little behind the input and does not rise instantly to its final value.

Now let us connect the feedback loop and let us assume that it has a completely flat frequency response. What we see in (B) is returned to the input, so that for this first trip the input will become something like the shape shown in (A) of Fig. 2 except that I have not

rounded off the corners. If we were to have this waveform we should obtain the output shown in (B) of Fig. 2. But if this were the output, the input would be the combination of this (multiplied by the feedback coefficient (β) and the input of Fig. 1 and would look something like the waveform shown in Fig. 3.

I have carried this exercise through to the point where it reveals the basic characteristics which are developing. We see the growth of that characteristic ring which we often observe in carrying out square-wave tests on feedback amplifiers and I do not think it unreasonable to guess that if we went round the loop a few more times the number of oscillations in the ring would increase in proportion. We know that in amplifier design we can modify this ringing characteristic, usually by increasing the feedback at high frequencies so that any tendency to hump upwards in the overall frequency characteristic is controlled.

We see also that the feedback does not have any effect on the amplitude of the signal applied at the leading edge of the transient. It is a matter of detail whether the system later moves monotonically or by an oscillatory path to the final state: the leading edge is unaffected. What is more, this is a necessary condition for the amplifier to be stable. I do not propose to go through this in any detail but you will find it in Wiener's "Cybernetics" (M.I.T. Press and John Wiley). The analysis is for a perfectly general system which delays the input $f(t)$ to give an output of $f(t - \tau)$. In this t is restricted to being less than the value corresponding to now because no one knows what the signal is going to be in the future. By considering a pulse signal which comes out after a first travel as $a_1 f(t - \tau)$ and as $a_k f(t - \tau)$ after k times round the loop we can progress rapidly to a general function of the form

$$\int_0^\infty a(t) f(t - \tau) dt$$

What Wiener does now is convert this to a frequency plane characteristic plot and thus reach, in the end, what we call Nyquist's criterion for stability. For a feedback amplifier to be stable it must have this characteristic of relying on the past and it must rely on the past in a rather restricted way. The way we

have treated the returned feedback signal deserves just a little mathematics. We put in some signal E and get out, first time round, just μE : we send back to the input $\beta\mu E$, so that we now get out the delayed signal $\mu \cdot \beta\mu E$. Back to the input goes $\beta\mu \cdot \beta\mu E$ to give a second delayed signal $\mu \cdot (\mu\beta)^2 E$. In fact the total output becomes:

$$\mu E (1 + \mu\beta + (\mu\beta)^2 + \dots)$$

This term in the parentheses is equal to $[1 - (\mu\beta)^n] / (1 - \mu\beta)$ and n is going towards infinity so that since $\mu\beta$ contains a delay term we get back to $\mu / (1 - \mu\beta)$ provided that we can forget about the unexploded bomb in the form of $(\mu\beta)^n$. If you want to go into the mathematics more deeply you will find that series like the one above are discussed in the standard textbooks, such as "Methods of Mathematical Physics," H. and B. S. Jeffreys, (Cambridge University Press, 1946). For our purposes, however, it is more than sufficient to notice that the classical feedback equation has got this hidden trap of time delay, without which the term $(\mu\beta)^n$ would be significant now and with which the leading edge of transients is uncontrolled by feedback.

The practical engineer may have become impatient with all this. He always shapes his feedback-path response so that there is no ringing when a square wave is applied. Surely that will be good enough for anyone. Unfortunately this is not so. Let us consider a rather simple design in which the output transformer has exactly the response we need for the whole amplifier, rolling off nicely at the highest frequency we wish to hear so that noise above this frequency will not be pumped out to cause intermodulation in the loud-speaker or a deflection on the customer's meter. The amplifier structure which precedes the transformer will then be designed to give a flat frequency response too and very low distortion. When we combine the amplifier and the transformer we get no spike on the leading edge and we believe all is well. But let us, just for the sake of having numbers, take the transformer cut-off frequency as 10,000 cps so that it will probably be 40 db down at 100,000 cps. If the amplifier produces a spike lasting 5 microseconds, which is a half cycle of 100,000 cps, and this spike is 10 times the normal signal amplitude we just will not see any trace of it after the transformer. Inside the amplifier, however, we have got this huge spike. Since we assume that the transient signal is one which drives the amplifier fully, the spike represents a fantastic overload.

Of course, it only lasts 5 microseconds and my poor old ears cannot de-

tect any intermodulation which is so short. I have raised before the question of whether we are designing equipment to entertain the local bats and for my part if the bats want music they can build and pay for the equipment themselves. The only trouble is that it isn't the fall which hurts, but the bruise. When the spike hits the grid of a tube near the end of the amplifier we shall drive a good deal of grid current through the grid-cathode diode and leave the coupling capacitor well charged. It will be a long time-constant circuit, because it is chosen to give good low-frequency response. The amplifier will become paralyzed for a time which is quite long enough to make itself known. We shall get the same result with transistors though we may, for circuit reasons, find that we cannot pump quite such a paralyzing charge into the capacitors.

The example taken is an artificial one and the reader may therefore be tempted to think that the problem will only appear in these artificial conditions and will not worry him when he is building a practical amplifier. In a transistor amplifier, however, we shall normally have a very restricted frequency range in the output transistors while the front end of the amplifier, using small transistors, will have a much more extended frequency response. The slow response of the output transistors will provide most of the delay in the forward amplification path but will prevent the spikes getting through to the output. The preamplifier will be overloaded, however, in just the way we have been discussing.

Even without a transformer we normally find that we have an even chance of designing our amplifiers to produce this effect. When we design a three-stage amplifier we know that if we want a reasonable amount of feedback we must design for either one narrow-band stage and two wide-band stages or one wide-band and two narrow-band stages. Nothing in the theory of stability gives us any reason why we should choose one of these arrangements rather than the other and nothing in the theory tells us why they should be in any particular order.

It is here, however, that our salvation lies. The input stage of an amplifier can usually handle a very much larger signal than will ever be applied to it and certainly it does not need very much extravagance in design to make sure that this is the case. If this stage is made to be a narrow-band stage it will prevent the passage further into the amplifier of the transient spike and will save the later stages from overloading. There is a good deal to be said in favor of making the input circuit, before the amplifier itself, restrict the bandwidth to the extent permitted. What is cer-

tainly needed is some way of making sure that in the interval before the signal has had time to get through the amplifier and back round the feedback loop the input cannot have risen to the point where overloading will take place.

An experimental study of this effect is not very easily carried out without fairly elaborate equipment. In tube circuits we might think that an oscilloscope connected at various points in the amplifier would do the job, but unless the oscilloscope has a very low capacitance it will disturb conditions in the circuit and may obscure the whole issue. In particular when we connect it in our wide-band circuits we may narrow the band-width to the point where the amplifier becomes unstable.

How important is this effect in the practical use of amplifiers? After all we rarely have to deal with transients representing anything like the full output. The practical situation is that we are using much of the grid bias for the low-frequency components of the sound and we must consider that the transient will be added to this. Since we are thinking in terms of factors of 10-30 times, corresponding to 20-30 db of feedback, realistic transient signals can make a firm bid for all the grid bias if added to the other signals can produce the overloading we fear.

This suggests that we might examine the amplifier by a sort of intermodulation test. If we apply a sustained low-frequency sinusoidal signal and a keyed high-frequency signal we can filter out the high frequency and see whether the low-frequency sine wave is at all disturbed by the sudden switching on of the high frequency. We must accept a keying click but must watch for a period of paralysis following the switching.

In the long run we will be trusting our ears. If it sounds all right it is all right and there is no point in carrying matters any further. If it does not sound right we start making all the measurements and the steady-state measurements will not reveal this effect at all. That is why it must not be forgotten. Transient distortion is difficult enough to measure, but transient intermodulation gives us the worst of two worlds: it is difficult to identify clearly by listening tests and difficult to simulate or measure. But oh, what an 'orrible noise! AE