Calculating Transient Response

Simplified Derivation from Frequency and Phase Characteristics

By THOMAS RODDAM

MACBETH, crossing a blasted heath, was not surprised to encounter three witches brewing up : nowadays he would expect to find a rather worn-looking mansion filled with engineers muttering : "When the mu-beta comes to one-nought, beware, beware!" There they sit, adding more and more feedback round more and more loops, and as the amplifiers get more linear, new troubles appear on the horizon. One of these is a problem which has long been discussed ; when you talk about 0.1 per cent. harmonic distortion, which harmonic? If you don't think it matters let me just remind you that you can hear a 1,000-c/s tone which is 36db lower than the quietest 100-c/s tone you can hear, so that 0.1 per cent. of 20th harmonic of 50 c/s is as audible as about 6 per cent. of second harmonic. As you know, feedback amplifiers tend to produce these higher harmonics in the overload region—in fact that is the only way you know that they are overloaded. The other problem is that of transient response.

There is an enormous amount of nonsense written about transient response, especially as it affects the loudspeaker. Broadcast programmes are amplified, piped round the country and re-amplified. In the course of this process, every effort is made to keep the frequency response flat, up to 10 kc/s, or whatever the frequency is. Above this, transformer after transformer provides a 12 db/octave cut-off. So what can you do, chum?

Transient response is important though for three, reasons: television, of course, in servo amplifiers and in amplifiers in tandem. The first two applications are fairly obvious, but the third deserves a fuller explanation. Suppose that we are operating two amplifiers, one driving the other: this is quite a common situation, for one may be a microphone amplifier and the other a power amplifier. If the first amplifier produces a large, though very short, overshoot there is no direct audible effect, because the frequencies in the transient peak are above the limit of hearing. This peak may, however, drive some stage of the power amplifier into grid current, and the stage may then be held at an improper bias by the grid CR network for an appreciable time. During this time, perhaps 1/10th of a second, the stage will produce more distortion than usual, and as the gain has been driven down, the feedback will be less effective than normal : muddy transients are the result.

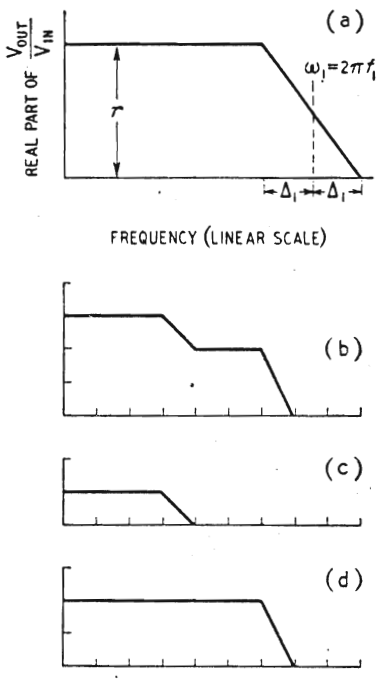


Fig. 1. (a) Basic trapezoidal response which gives the transient response shown in Fig. 2. The more complicated response in (b) may be resolved into the two trapezoidal responses shown in (c) and (d).

It need not be two amplifiers in two boxes for this effect to be apparent. A single multistage amplifier with a transformer in the middle, or perhaps with the feedback arranged in two separate loops can cause trouble of this kind. As our amplifier designs get more and more sophisticated we need to watch out for more and more of these obscure effects.

Circumventing Laplace

The obvious thing for the conscientious designer to do is to calculate the transient response, just as he calculates the frequency response of a feedback amplifier before he starts. Very few designers do this, because they imagine they will be confronted by an immense formula to be fed into the Laplace Transform machine. If this were the only way of studying transient response they would—to use an Antipodean phrase—be too right. Fortunately G. F. Floyd, of the Massachussetts Institute of Technology, has described in a thesis* a simpler way of dealing with the problem. Floyd's method, dehydrated and predigested and generally made fit for engineers' consumption, is the subject to reason about to-day.

First of all we need to know the frequency response and phase characteristic of the amplifier. In all the discussion which follows I propose to treat only the transient due to the high-frequency cut-off, and not the droop due to lack of low-frequency response. I shall, however, comment on the application of the method to "droop" calculation at the end of the article. But back to our frequency response. If you have read any of the papers or books† on the connection between response in time and frequency response you may have noticed that the decibel and the logarithmic frequency scale are not used. We use these logarithmic units for convenience, and because our ears are fairly logarithmic in performance. When considering transient response we start off with a very

* See also "Principles of Servomechanisms" by G. S. Brown and D. P. Campbell, published by Chapman & Hall.

† For example, "Radio Engineers' Handbook" by F. E. Terman, published by McGraw Hill, p. 968 et seq.

artificial signal and examine how the energy it contains gets redistributed in time, and the energy in one frequency band is as important as that in the neighbouring bands. We must therefore be prepared to draw our frequency response in terms of actual magnification with a linear frequency scale.

There are two separate meanings attached to the term "transient response." To some people it means the response to a square-wave input, while to others it means the response to a very short impulse. Between the two there is, of course, a very close mathematical connection, and it is largely a matter of practical convenience which is used. I prefer to use a short impulse for test purposes, because it is then so easy to recognize the unwanted "ringing" which appears as a train of damped oscillations after the main impulse. In some special circumstances, however, square waves provide much more information: for example, I have arranged amplifiers so that they were unstable on the "swing" of a square wave and stable on the "swong." On the oscilloscope the unstable condition was shown by a growing oscillation on one half of the cycle, followed by a decaying oscillation on the other half. This sort of behaviour corresponds to those bursts of high-frequency oscillation at the low-frequency peaks, which produce such an unpleasant sound and are rather difficult to detect without very full tests.

For the purpose of this article, transient response will be taken as the response to a very short impulse. The method of calculating it depends on a basic theorem, the truth of which we assume in almost all our electrical theory. That theorem is the Superposition Theorem, which states, though not in these words, that two happenings in a linear system go on quite independently of each other. To calculate the transient response we first of all imagine a circuit having a particular frequency response for which the transient response is easily calculated. We then pretend that the actual circuit is made up of a number of these ideal systems in parallel, having different factors in their make-up. Each system passes a transient of the standard type of a particular size and time scale. Then we add all the transient voltages at any instant together. An example will make this clearer.

The standard frequency response, which is called, for reasons which will follow, the "real part" response, is shown in Fig. 1(a). The transient response of an amplifier (or any other network) having this frequency response is shown in Fig.2. Suppose that we find that our amplifier has a response like that shown in Fig. 1(b). We imagine it to be made up from two units, one having the response of Fig. 1(c) and one having the response of Fig. 1(d). We take two curves of the form of Fig. 2 with the appropriate scales, add them, and there is the final transient response of the system.

"Real Part" Response

We must now begin to clothe these bare bones. As I said above, we use a frequency response known as the real part response. This is the graph of $m \cos \theta$, where m is the voltage gain and θ the phase shift between input and output. For calculation purposes this is a great nuisance, because if you are using graphical methods of predicting the frequency response you are working with decibels and a logarithmic frequency scale.

The stages in the determination of the real part response can be followed in Fig. 3. The basic response

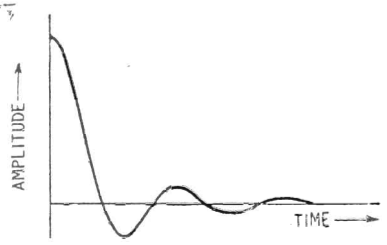


Fig. 2. Response to a single short impulse of a circuit with the frequency response of Fig. 1 (a).

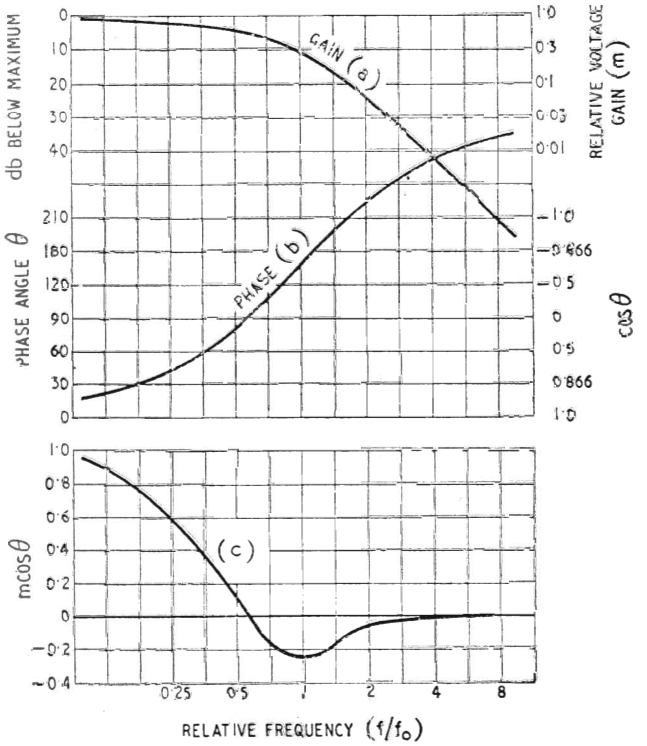


Fig. 3. From the gain and phase to characteristics (a) and (b), the real part response (c) is calculated. Frequencies are plotted on an octave (log) scale.

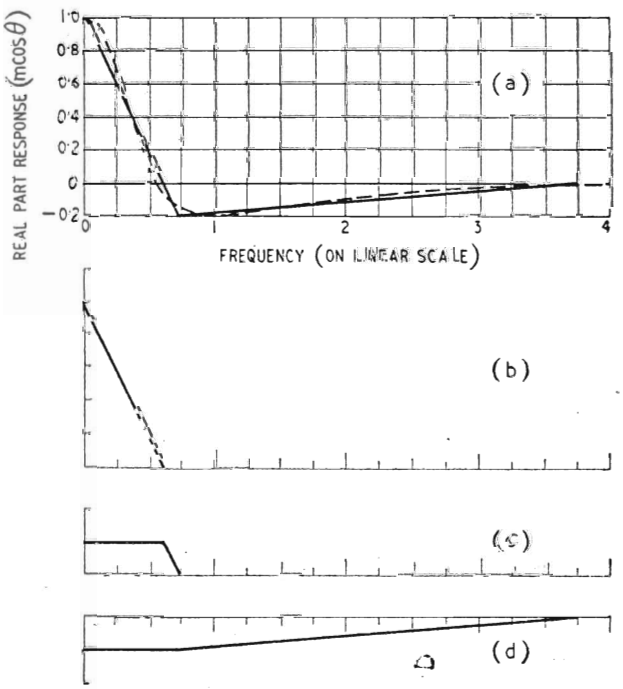


Fig. 4. (a) Real part response of Fig. 3 (c) replotted to a linear frequency scale with superimposed straight-line approximation. (b) (c) and (d) Trapezoidal responses which added together give the straight line approximation of (a).

curves, frequency response in decibels, and the phase characteristic, are shown at (a) and (b). The frequency scale here is an octave scale, which for all theoretical purposes is just as good as a logarithmic scale of the ordinary kind, in fact you can mark a logarithmic scale in on the paper. For practical purposes, however, it is very much better, because it uses ordinary centimetre square paper, which is cheaper and is always available. It is painfully surprising how often in a large organization the 3-decade log paper runs out, and you have to make do with 2-decade or 4-decade. And if you have standard curve shapes, they just don't fit the paper.

In addition to the decibel and angle scales I have marked in the voltage ratio, m, and cos θ scales. At each convenient frequency we take (voltage ratio) $\times$ (cos θ) to get the real part (RP) response plotted in Fig. 3(c). This RP response is still plotted on an octave frequency scale, and it must be replotted on a linear frequency scale before we can use it. This has been done in Fig. 4(a), which shows clearly how the logarithmic frequency scale tends to minimize the very important high-frequency behaviour.

The solid line segment response shown in Fig. 4(a) is the approximate form which is used for calculating the transient response. It is not too difficult to see that this can be represented as the sum of the three RP responses shown as Figs. 4(b), (c), (d), which are all of the standard trapezoidal form. All we need to do now is to take the transient response corresponding to each trapezium and add them together (that for (d) of course, must be subtracted).

At this point we introduce the essential formula. With the terms defined in Fig. 1(a), the transient response of a system having a trapezoidal real part characteristic is given by the equation

$$h(t) = \frac{2r}{\pi} \cdot \omega_1 \left(\frac{\sin\ \omega_1 t}{\omega_1 t}\right) \left(\frac{\sin \Delta_1 t}{\Delta_1 t}\right)$$

whatever you do about it, this formula involves quite a lot of arithmetic. The linear is simplified by making use of a table or graph of the function (sin x/x). I have produced a graph of this function, and it is given at Fig. 5. For each trapezium we then make a table of the form shown:

TABLE

Trapezium 1.:

$r =$	$\omega_1 =$
$\frac{2r}{\pi} =$	$\Delta_1 =$

t	$\omega_1 t$	$\Delta_1 t$	$\frac{\sin \omega_1 t}{\omega_1 t}$	$\frac{\sin \Delta_1 t}{\Delta_1 t}$	$h(t)$
0			from Fig. 5		
1/10,000					
2/10,000					
etc.					

We then transfer the last column to a new table, in which the transient responses for the separate trapezium are collected. Adding the response for each time we have the total transient response: $h(t) = h_1(t) + h_2(t) + \ldots\ldots\ldots$

Kronecker, who introduced the delta function, a kind of unit impulse, into analysis, says somewhere: "God made the integer; the rest is the work of man." He could have hardly been more right about this particular impulse problem, because the average network transient response takes about a page full of closely written figures. But it is only slide rule work and addition, there is no real mathematics to it. I have not carried through the calculation of the transient response for one example: it would make an impressive looking page, but I do not think the Editor really likes to publish a page of dull arithmetic.

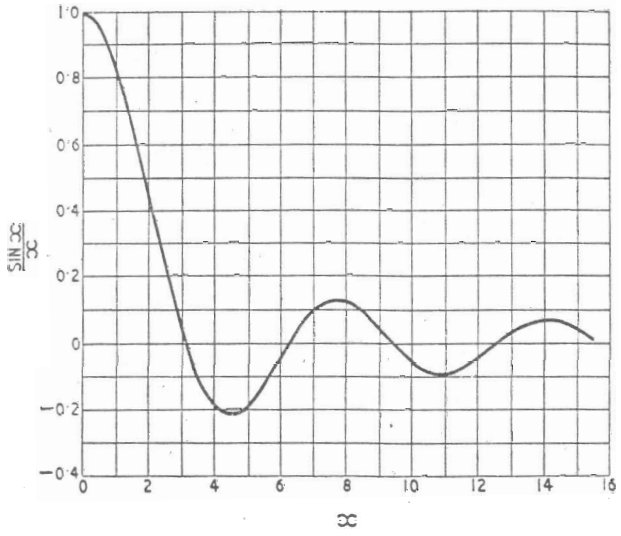


Fig. 5. Graph of the function sin x/x.

Before going on to some related topics let me recapitulate. To find the response to a very short pulse, we find the frequency and phase characteristics, compute the real part response by taking (gain) $\times$ cos (phase) at each frequency, plot this on a linear frequency scale. This real part response is then expressed as a set of trapezoidal responses and the transient response corresponding to each trapezium is calculated. Finally, we add (or subtract where necessary) all these transient responses and this is the transient response of the complete system.

When you are considering overload problems, the square-wave response is probably more convenient than the impulse response. The overshoot of the square wave gives a pretty good idea of how much safety margin must be allowed between the steady-state output and the maximum programme output. Fortunately, it is very easy to proceed from impulse response to square-wave response. In Fig. 6(a) I have drawn a fairly typical impulse response. The figures inside the curve show the number of millimetre squares in each 1-cm vertical strip, and the running total from left to right is underneath. This running total is then plotted in Fig. 6(b) and shows the square-wave response corresponding to the given impulse response. The overshoot is very small, so that even the most cautious user of an amplifier with this characteristic could operate it up to its steady tone maximum. There is nothing more to the calculation of the square-wave response: all the work is in the first stage, the determination of the impulse response.

"Ringing"

It is very useful to have some physical appreciation of the causes of the ripples in the transient response. The simplest view is obtained by noticing that our input signal contains a complete frequency spectrum extending up towards infinity. The amplifier cut-off stops all frequencies above a particular limit and therefore acts as though it produced a negative signal containing all these components with phase reversed to add to the original signal. Some of the filter text-books show what happens to a square wave passed through a high-pass filter, and this damped oscillation is the wave to be subtracted from a square wave which has passed through a low-pass circuit. The

reader who checks up on this will no doubt ask why I haven't pointed out that the filter books also show the transient response of a low-pass circuit. The reason is that I wish to emphasize the fact that the transient distortion is due to terms above the maximum transmitted frequency and that they are not in themselves audible.

Feedback amplifiers present some special and rather interesting transient characteristics. To the mathematician the reason is very simple : the phase characteristic hugs the zero line up to near the cut-off, and then rises very sharply. That's fine, but what does it mean ?

Delayed Feedback

The easy way to understand what happens in a feedback amplifier is to watch a pulse going through it. A typical amplifier, let us say, has a frequency response without feedback which is 20 db down at 20 kc/s, and we are using 20 db of feedback. The phase shift will probably be 180 degrees at about 25 kc/s. The delay through the amplifier, without feedback, is given by the shape of the phase characteristic, $d\theta/d\omega$, and the average value of this is $\pi/2\pi.25{,}000$ or $1/50{,}000$ sec, or 20 μsec later. Until the pulse reaches the output the feedback cannot begin to have any effect, so that with a square-wave input the first 20 μsec of output is amplified by the full gain of the amplifier. At the end of 20 μsec the feedback starts to operate, but during this short period you may have blocked off a grid somewhere in the circuit. I do not pretend that the description here is complete : it is, however, of very great value if you are designing an amplifier with feedback round an output transformer, when the early stages are usually made with very wide band response in order to achieve stability. These conditions lead to a pulse at the output grid which may be about ten times the size of the steady-state signal. There are quite a lot of complications which can arise in particular circumstances, but I do not propose to discuss them here.

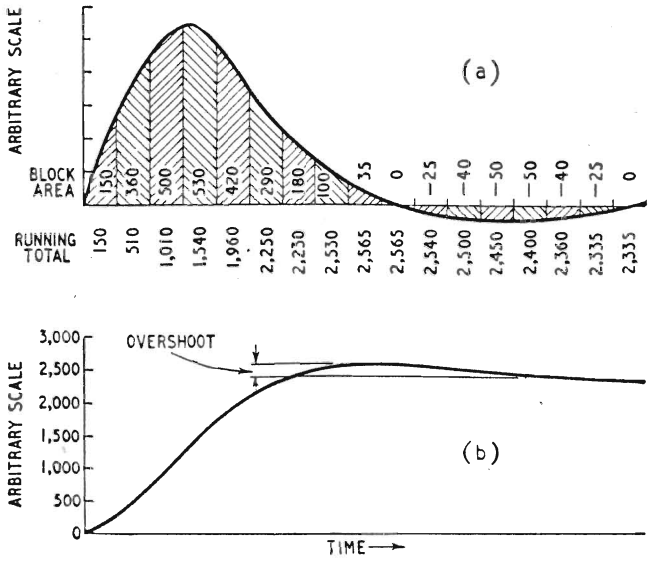
It is not suggested that in all cases you should calculate the transient response before you build an amplifier. One feature of the theoretical method, however, is that it provides a background which helps in interpreting the transient responses you can see on the oscilloscope.

Square-wave responses of amplifiers have often been published in *Wireless World*, but there is a method of studying the transient response which has not been mentioned, so far as I can remember. This is to differentiate the amplifier output and thus obtain the impulse response, at the same time making sure that the amplifier behaves well with both negative and positive swings.

The differentiating circuit is simple, a series capacitance and shunt resistance after the load and before the oscilloscope. The shunt resistance should be fairly large compared with the load, while the capacitance should be chosen to give high impedance compared with the shunt resistance at all frequencies of interest. Typical values would be 100 pF and a few thousand ohms. The advantage of this method is that it shows up the ripples on the response much more clearly : ideally the square wave when differentiated will just give a spike of very short duration ; all else is error.

To conclude, a note on the calculation of " droop " caused by a bad low-frequency response is needed. The procedure here is fairly simple : you plot the real part response in just the usual way, and then take the trapezium, or set of trapezia, which would be needed to make the response go down to zero frequency. The impulse response is then calculated for these trapezia, and the square-wave response obtained by counting squares (integration). This response is subtracted from the ideal square wave, and there is your drooping characteristic.

Fig. 6. Square-wave response is obtained by counting squares in the elements of area under the impulse response curve, and then plotting the running total. (a) Typical impulse response. (b) Corresponding square-wave response.



R.I.C. SPECIFICATIONS

THREE new component specifications and additional sections for some existing ones have just been issued by the Radio Industry Council. These specifications are prepared in conjunction with B.R.E.M.A., R.C.E.E.A. and R.E.C.M.F. and are for the time being intended for use within the industry.

Sections 1 and 2 of the new specifications are available now and these cover performance requirements and production tests. Sections 3 of each, defining types of the components covered, their ratings and sizes, will be issued later.

RIC/151 deals with dolly operated switches of the toggle type for use in d.c. and a.c. circuits not exceeding 500 V and 15 A loading and for frequencies up to 3 kc/s. RIC/154 relates to single- and multi-wafer rotary switches and concerns two types; one for use up to 100 kc/s and the other up to 100 Mc/s.

RIC/251 deals with valveholders of the kind commonly used in radio receivers and other electronic equipment. It covers two types of valveholders; those with low loss insulating material having power factors below 0.002 at 1 Mc/s and those with poorer material with power factors greater than 0.002 at 1 Mc/s. RIC/151 and 154 cost 6s each and RIC/251 5s 6d; this is inclusive of part 3 in each case, which will be supplied later.

The additional sections now available are parts 3 for RIC/111, non-insulated wire-wound resistors, and defines the standard values, tolerances, sizes and finishes; for RIC/122, variable-track composition resistors, again giving values and also switch ratings when fitted; and for RIC/133, defining values, tolerances and ratings of fixed ceramic grade 1 capacitors. These complete the three specifications concerned.