See note at the end of the text

SECTION 2—*Instrumentation and Methods*

74 Properties and Limitations of Electronic Instrumentation

JAN-ERIK SIGDELL

THE PURPOSE OF THIS CHAPTER is to describe certain properties and limitations of electronic instrumentation to the user who is not an expert in electronics and who may have relatively little theoretic knowledge of electronics in general. Mathematical exposés are reduced to a minimum; when used, they are accompanied by verbal explanations. Many graphic demonstrations are used. Still, some mathematics cannot be avoided, although the aim is rather a phenomenologic description.

This chapter deals with "amplifiers," but in a general and broad sense. The term "amplifier" here means almost any linear electronic device with an input and an output. It also may be a filter or an attenuator (power *amplification* less than unity means power *attenuation*) and may even include a recorder (the trace on a paper as output) or a transducer (e.g., mechanical input).

Frequency Characteristics of Broad-Band Amplifiers

The term broad-band amplifier will be used to characterize a linear amplifier having constant amplification within a certain frequency range, which is not small as compared to the individual frequencies within it. The narrow-band amplifier, in which the band of constant amplification extends between two closely spaced frequencies, will be discussed later.

First, the concept of amplification. By this, we mean the output and input quotient between sinusoidal quantities. Limiting to the case in which both these quantities are voltages, this is:

$$A(f) = \frac{e_o(f)}{e_i(f)} \qquad (74\text{-}1)$$

where $A(f)$ is the amplification and $e_o(f)$ the sinusoidal output voltage, corresponding to a sinusoidal input voltage $e_i(f)$, each taken as functions of frequency. These quantities are given in *complex notation*. In this notation, a sinusoidal signal $E\sin(2\pi ft + \varphi)$ is written:

$$e = Ee^{j\varphi} = E\cos\varphi + jE\sin\varphi \qquad (74\text{-}2)^*$$

where $j = \sqrt{-1}$, E = amplitude and φ = phase. This makes it natural to distinguish amplitude $|A|$ and phase $\angle A$ of the amplification:

$$A = |A|e^{j\angle A} \qquad (74\text{-}3)^\dagger$$

where $|A|$ and $\angle A$ vary with frequency.

The term "amplification" can, of course, be generalized to any input and output quantity, but A is dimensionless only when those quantities are both of the same dimension, as above. The following discussion holds for most linear electronic devices, even if input and output are not voltages.

As a law of nature, the range of constant amplification cannot extend to infinite frequency; often the range is intentionally limited to a determined range, as is almost always the case in biomedical applications—except for recorders, for example, where the range extends to the natural limitations of the system; this range usually is made entire use of and sometimes, even then, it still is insufficient. Amplification is thus gradually reduced as frequency increases above a predetermined range, limited intentionally or naturally. This gradual decrease of amplification is called *roll-off* and can be shown to amount asymptotically to 6 n db/octave, or 20 n db/decade, where n is an integer for discrete component amplifiers. An octave is the range from one frequency to its double value; a decade is the range to ten times an initial value. If the range of constant amplification does not extend down to d-c (zero frequency) an analogous low-frequency roll-off will occur, with amplification falling with decreasing frequency, also following a 6 n db/octave asymptotic behavior. Actually, amplification can never be exactly constant over any range, but its variation can be kept small within the main part of a determined range.

As a measure of range of "constant" amplification, one generally chooses the *3-db-bandwidth*. This concept will be used throughout this chapter and will be referred to simply as *bandwidth* (in some technical applications, one also uses the 6 db-bandwidth). This bandwidth is the difference between the *upper* and *lower cutoff frequencies* at points at which A has fallen 3 db, i.e., by a factor of $1/\sqrt{2} \approx 0.707$, below its mid-range (maximal) value. The corresponding frequency range is called a *passband*. The upper and lower cutoff frequencies sometimes are called "half power frequencies," as a 3-db reduction of the output voltage corresponds to halving the power in a constant load.

The frequency characteristic of an amplifier may exhibit "constant" levels within several bands. In such cases, generally only the band with the highest amplification is used as a passband. The amplifier characteristic in terms of the variation of $|A|$ with frequency often is plotted in a log-log diagram (or, equivalently, a lin-log or semilog diagram if $|A|$ is expressed in db on the linear axis, $|A|$ in db is $20_{10}\log|A|$).

The asymptotic behavior mentioned above means that this characteristic can be approximated by a polygonal combination of straight lines (piecewise linear approximation), being horizontal or sloping ± 6 n db/octave. Such a representation is called a *Bode diagram*. A representative example is shown in Figure 74-1, where the straight lines are continuously drawn and the exact amplitude characteristic given by the dashed curve. In Figure 74-1, the upper

*This actually is the definition of the exponential function with imaginary exponent: "e raised to $j\varphi$ plus j times sine of φ." e = 2.718 281 828 459 045 23 . . .. the base of the natural logarithm.

†With properly varying "φ," this is an analogy to equation (72-2), expressing A as the magnitude of A times the exponential of the product of j with phase of A.

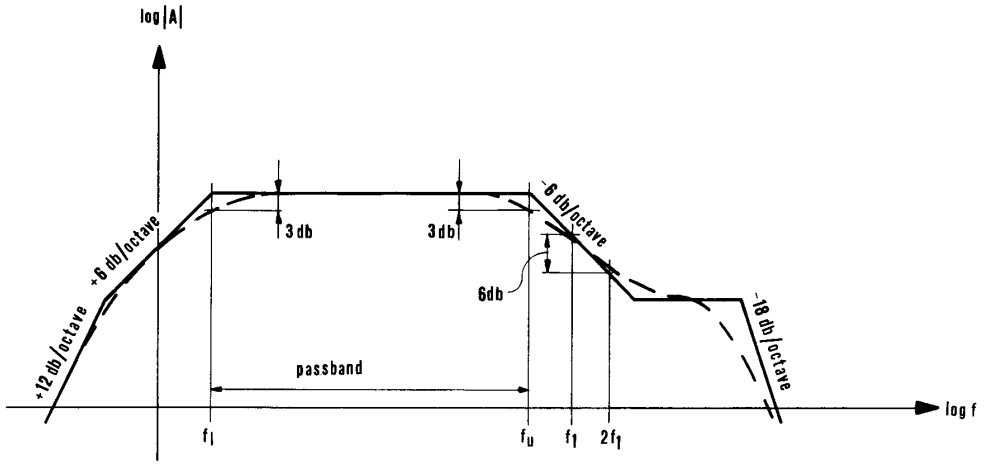


Fig. 74-1.—An example of the characteristic of an amplifier. Solid line: asymptotic characteristic; dashed curve: true characteristic.

and lower cutoff frequencies are marked by f_u and f_l, respectively. Hence, the bandwidth is $B = f_u - f_l$. The points at which the horizontal passband line in the asymptotic representation meets the upper and lower roll-off asymptotes are called *corner frequencies*. If the asymptotic roll-offs are 6 db/octave on both sides of the passband within sufficiently broad ranges, corner and cutoff frequencies coincide. If the asymptotic roll-offs start at 6 n db/octave, the cutoff frequencies will be related to the upper corner frequency f_{cu} and the lower corner frequency f_{cl} as

$$f_u = f_{cu}\sqrt{2^{1/n} - 1} \qquad (74\text{-}4)^*$$

and

$$f_l = \frac{f_{cl}}{\sqrt{2^{1/n} - 1}} \qquad (74\text{-}5)^\dagger$$

where n may be different for f_u and f_l.

Operation Near or Outside the Cutoff Frequencies

As long as the sinusoidal signal remains well within the passband, the output corresponding to a constant-amplitude input also remains at constant amplitude. As a cutoff frequency is approached, the output amplitude is gradually reduced to a minimal value (at the cutoff frequency) of $1/\sqrt{2}$ of the mid-passband amplitude. As the frequency is shifted to outside the passband, output amplitude falls off faster according to the roll-off. Therefore, a change in amplitude of the output signal is necessarily caused only by a proportional amplitude change in the input signal as long as the signal remains well within the passband. Near the ends or on the roll-off, the change could also have been caused—partially or totally—by a shift in the frequency.

For some purposes, such as for the improvement of phase behavior, one may raise the amplitude (within limits, e.g., ±3 db variation in the passband) somewhat, near the cutoff frequencies, before reaching the roll-off. If the amplitude thus increases first before going down 3 db, the in-passband variation of amplitude with frequency is even greater, near the cutoff frequencies. (Compare to the description of second order systems in Chapter 15).

*The steeper the roll-off the lower the f_u for a given f_{cu}, as "f_u is f_{cu} times the square root of the n:th root of 2, minus 1".

†The steeper the roll-off the higher the f_l for a given f_{cl}.

Phase Considerations

Nearly constant amplification within the passband should also be accompanied by a fairly constant phase shift $\angle A$ between input and output. For most amplifiers, this phase shift can be shown to be practically zero around the mid-passband. (Networks for special filtering purposes, having other, generally varying, phase shift within a band of constant $\angle A$, may be constructed but will not be considered here.) The variation of the phase $\angle A$ near the cutoff frequencies generally is larger than the amplitude variation. As an example, the characteristic of an RC-coupled amplifier with −6 db/octave asymptotic roll-off in a broad range above the upper cutoff frequency, f_u, is sketched in Figure 74-2, in a region around f_u. The behavior is analogous at the lower cutoff frequency for a 6 db/octave asymptotic low-frequency cutoff.

The phase $\angle A$ can be considered constant only within a central part of the passband (or for frequencies $\ll f_u$ if $f_l = 0$), much smaller than the bandwidth, for an amplifier with a characteristic as sketched in Figure 74-2. If the asymptotic roll-off above f_u (or below f_l) is faster than 6 db/octave near f_u, the phase shift A of the signal varies even more within the passband. With special circuits, the situation can be improved somewhat, but it is much easier to build amplifiers with non-constant but approximately linear phase characteristics within their passbands. Such a characteristic is of considerable importance for low-distortion amplification, such as for aperiodic signals, discussed below.

These phase considerations show that comparisons between signals of the same frequency, amplified in different amplifiers, can involve large errors if the amplifiers are not identical. If phase shift comparisons are to be made between two signals, a given frequency and one of its multiples, this generally can be done with good accuracy only if both signals fall much nearer the center than the ends of the passband. One can, of course, correct for errors due to phase shift within the amplifier if its phase characteristics are clearly known.

Periodic Signals of Non-Sinusoidal Waveform

As is well known, periodic non-sinusoidal signals can be represented by or synthesized from a series of harmonically

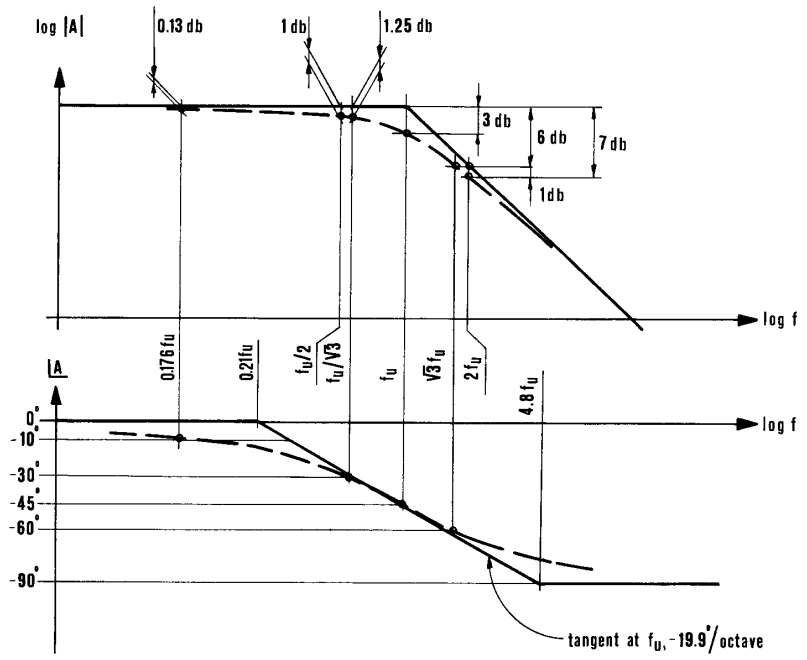


Fig. 74-2.—The characteristic around the upper cutoff frequency f_u for an RC-coupled amplifier stage. Solid lines: asymptotic characteristic; dashed curves: true characteristic.

related sinusoidal components (i.e., having frequencies that are multiples of a basic frequency), a *Fourier series*:

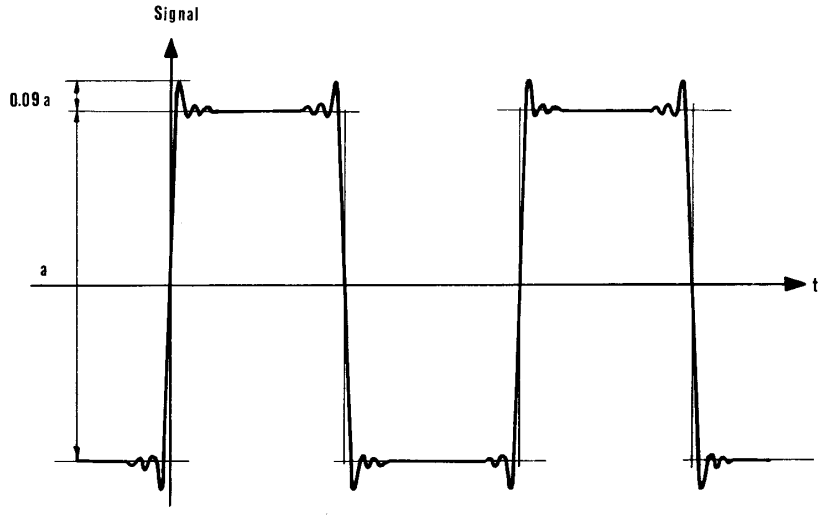
$$s(t) = \sum_{k=0}^{\infty} [a_k \sin(2k\pi f_0 t) + b_k \cos(2k\pi f_0 t)] \quad (74\text{-}6)^*$$

where f_0 is the (basic) frequency of the periodic signal $s(t)$. When amplifying such signals, it is, of course, important that all relevant components fall well within the passband of the amplifier. The relevant components are the first N components, which are needed for a good approximation, as

$$s(t) \approx \sum_{k=0}^{N} [a_k \sin(2k\pi f_0 t) + b_k \cos(2k\pi f_0 t)] \quad (74\text{-}7)^\dagger$$

Because of the intentions of this chapter, mathematical

Fig. 74-3.—Gibbs' phenomenon, caused by truncation of the Fourier series for a square wave.



*Any (practical) periodic time function can be synthesized from a (generally infinite) number of sinusoidal functions with proper amplitudes (and phases). "s, function of t, is the sum, from frequency zero to frequency infinity, of linear combinations of a sine and a cosine of the same frequency, all frequencies being integer multiples of f_o".

†The higher-numbered components in equation (74-6) have less influence and may be ignored from a certain number on with only a minor error in the representation of $s(t)$. "$s(t)$ is approximately equal to the sum from $k = 0$ to $k = N$. . ."

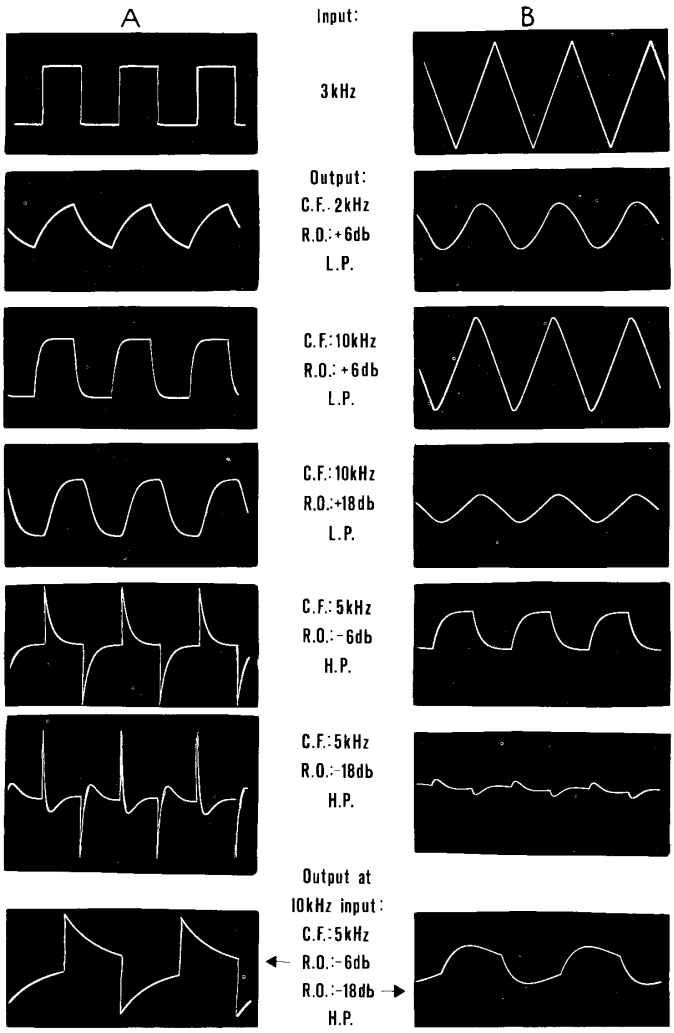


Fig. 74-4. – Influence of the amplifier response on signals (RC-coupled stages). In **A**: for a square wave; in **B**: for a triangular wave. C.F. = corner frequency, R.O. = roll-off, L.P. = low-pass (influence of a high-frequency roll-off), H.P. = high-pass (influence of a low-frequency roll-off). (Photographed from oscillograph tracings.)

reasoning will be kept to a minimum. Therefore, the influence of the amplitude and phase characteristics on a periodic non-sinusoidal signal will be discussed only briefly in the text, being further illustrated by some examples.

A sharp cutoff in the amplitude characteristic will more or less truncate the Fourier series for the periodic signal. If this occurs well above Nf_o, where N is defined by equation (74-7), the influence of the amplitude characteristic will be negligible. But if it occurs at a lower frequency, the signal may be strongly distorted. The distorting effect is higher in the vicinity of the fast transition parts of the input signal. In the extreme case of a square wave, a pure truncation (infinitely sharp cutoff) of the corresponding Fourier series leads to the appearance of *Gibbs' phenomenon* in the output signal, as illustrated in Figure 74-3. This phenomenon is related to *ringing* in an amplifier.

Normal amplifiers generally will not show the extremes as sketched in Figure 74-3; also, "ringing" cannot occur *before* the transition with any realizable electronic amplifer, since it cannot have an infinitely sharp cutoff. Nevertheless, it should be kept in mind that oscillations observed *after* fast transitions in the signal could be caused by "ringing" in the amplifier. Another effect is a *smoothing* of the transition, which will be discussed further in relation to rise-time. Clearly, peaks and sharp "edges" in the signal may be obscured by such "ringing" or more or less erased by a smoothing effect. As is discussed further below, details that occur during short or fast parts of the signal are related to higher frequencies in the Fourier series. This may also be understood from equation (74-6). Frequency components considerably lower than $1/t_o$ cannot contribute significantly to details in the signal occurring during shorter times than t_o, as the sinusoids associated with each of those components vary too slowly. As an example, a bandwidth of the order of a few hundred Hz may be needed to be able to observe fine details in the ECG, such as small changes in the Q, R or S peaks. A narrow bandwidth will always round off the peaks and may obscure or erase details of peaks or steep slopes, although the main form of the QRS complex may be well preserved.

The influence of a phase shift on a signal component in equation (74-6) is a translation of the component along the time scale, having the effect on the signal of adding the difference between the translated component and the true

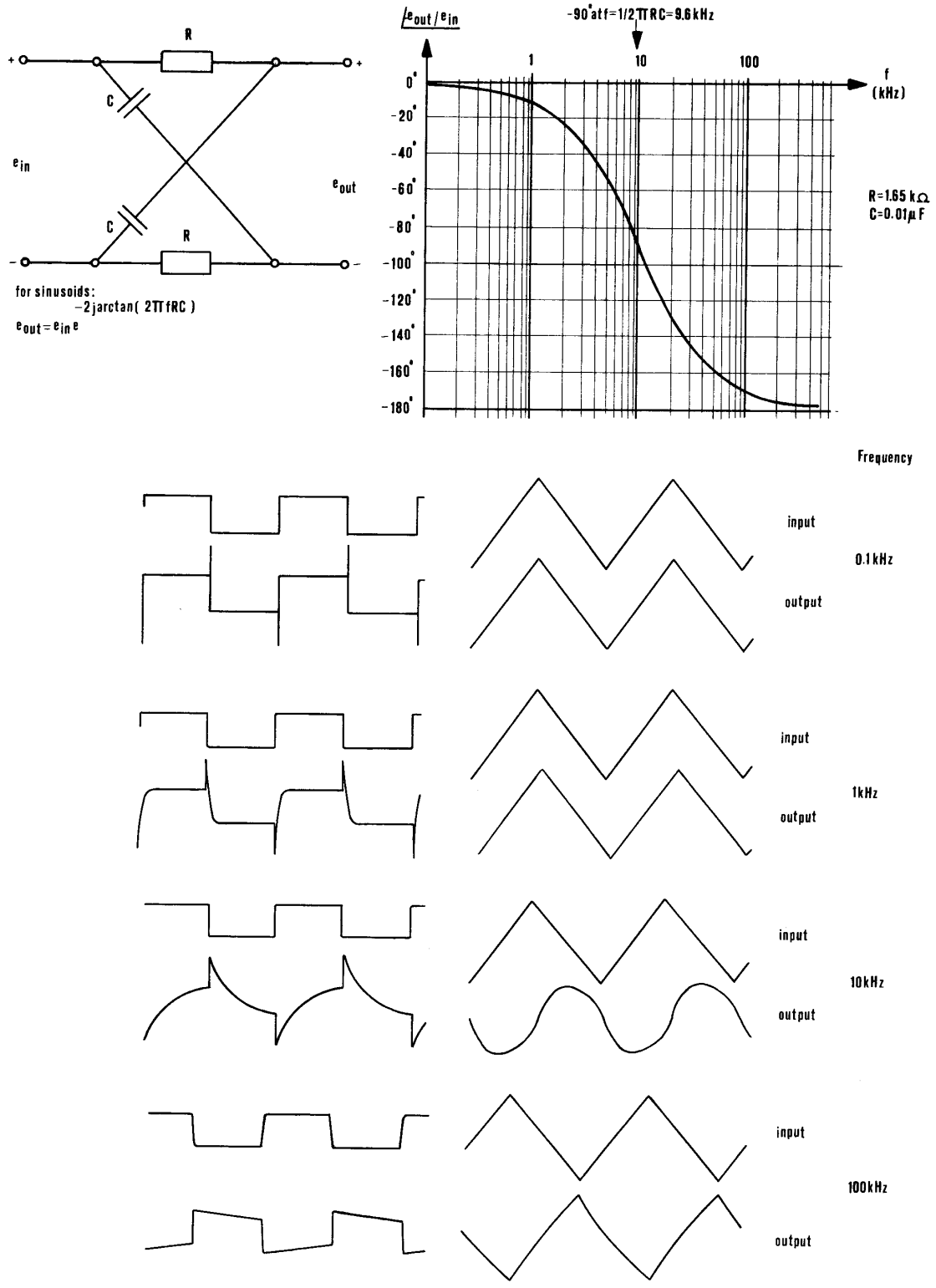


Fig. 74-5.—Influence of a mere phase shift on signals. The sketched circuit (all-pass filter), with the phase characteristic shown in the diagram, causes the distortions shown at different frequencies. (Redrawn from photographed oscillographic recordings.)

component (both of which are sinusoidal time functions of the same frequency). Damping or amplification of a single component produces a similar effect. When all components together are subject to varying phase shifts, such effects may average out to a mere delay of the entire signal without distortion of its form. This occurs when the phase shift varies *linearly* with frequency within the important frequency range, as will be discussed further below (a delay of time τ of the signal appears, conversely, as a phase shift $2k\pi f_0\tau$ the k:th component in equation [74-6] as is easily seen).

Similarly, a low-frequency roll-off at or above the basic frequency has an effect comparable to subtracting sinusoids of the frequencies falling in the roll-off region (with less amplitude and phase shift as these frequencies approach the lower limit of the passband).

The nature of such influences is best demonstrated by a few examples. Figures 74-4 and 74-5 show the influences of amplitude and phase characteristics on square and triangular waves. Due to practical difficulties in producing these oscilloscope tracings, Figure 74-4 does not show the influence of variations in amplitude response alone, but coupled with an associated variation in the phase response (see Fig. 74-2). Figure 74-5, on the other hand, shows the influence of a mere phase shift, realized with an "all-pass filter" (having constant amplitude $|A|$ but varying phase $\underline{/A}$ within a broad range).

Cascaded Amplifiers

If amplifier stages are connected together in a series arrangement, so that the output of one is the input to the next, they are said to be connected in cascade. The total amplification of a series of cascaded amplifiers is the product of all the individual amplifications. Here, one must generally consider the effect on the amplification of a single stage, caused by the loading (input impedance) of the following stage and the source (output) impedance of the preceding stage; see the discussion on input and output characteristics in a later part of this chapter. The above applies to the cascaded amplifiers considered as a whole, i.e., with the mentioned, total amplification.

If amplifier stages are cascaded, having *identical* amplifications A (considering the effects of input and output impedances), the total amplification is

$$A_{\text{tot}} = A^n \tag{74-8}$$

The asymptotic roll-off slopes in the Bode diagram of the total amplification is n times larger than for A. If the latter slope is 6 m db/octave, the former is 6 mn db/octave near the cutoffs, and we can find the relation between upper and lower cutoff frequencies for A_{tot} and A, using equations (74-4) and (74-5). Eliminating corner frequencies, it follows:

$$f_{u\text{tot}} = f_u k, \tag{74-9}*$$

$$f_{l\text{tot}} = \frac{f_l}{k} \tag{74-10}*$$

where f_u and f_l are the upper and lower cutoff frequencies of the individual stages and

$$k = \sqrt{\frac{2^{1/mn} - 1}{2^{1/m} - 1}} \tag{74-11}†$$

where m may be different for f_u and f_l. In the common case that $f_u \gg f_l$, we find for the bandwidth of A_{tot}:

$$B_{\text{tot}} \approx Bk \tag{74-12}$$

where B is the bandwidth of A. The more general case is derived from equations (74-9) and (74-10).

Time Response of Broad-Band Amplifiers

Whereas above we considered the influence of the amplifier (its amplitude and phase characteristics) on periodic signals, here we will consider the influences on aperiodic signals, such as pulses, evoked responses, etc.

Time-Frequency Relations

Events in a periodic signal, which occur rapidly or extend over short times, are related to the higher-frequency components in the Fourier series for that periodic signal. This holds just as well for an aperiodic signal. For such signals, the Fourier series (equation[6]) can be generalized to a Fourier integral:

$$s(t) = \int_{-\infty}^{\infty} S(f)e^{j2\pi ft}\,df \tag{74-13}*$$

where $S(f)$ is the spectrum of the signal, obtained as

$$S(f) = \int_{-\infty}^{\infty} s(t)e^{-j2\pi ft}dt \tag{74-14}†$$

We no longer are dealing with a sequence of discrete frequency components, but with a continuum, a frequency distribution, a spectrum. A spectrum generally is complex; i.e., it can be separated in an amplitude spectrum $|S(f)|$ and a phase spectrum $\underline{/S(f)}$:

$$S(f) = |S(f)|\,e^{j\,\underline{/S(f)}} \tag{74-15}‡$$

The relation between input and output through the amplification $A(f)$ is

$$S_o(f) = A(f) \cdot S_i(f) \tag{74-16}$$

where S_o and S_i are the spectra for the output and input signals, respectively. This is a generalization of equation (74-1); in this context, $A(f)$ usually is referred to as a *transfer function*, since it relates spectra and not signals directly (except in the case of a pure sinusoid). Still, $A(f)$ has the same function as in equation (74-1). Equation (74-14) can be written

$$S(f) = \int_{-\infty}^{\infty} s(t)\cos(2\pi ft)dt - j\int_{-\infty}^{\infty} s(t)\sin(2\pi ft)dt \tag{74-17}§$$

From this, we may see intuitively that a shorter time duration of the signal corresponds to a higher-frequency content in the spectrum. The sinusoidal oscillations of the trigonometric functions in the integrals of equation (74-17) occur at a faster rate the higher the frequency f. These oscillations tend to average out at integration, especially where there are more oscillations within the duration of $s(t)$. Thus, a fall-off of $|S(f)|$ with increasing frequency generally is expected and will occur faster and at lower frequencies the longer the duration of $s(t)$ (many oscillations of the trigonometric functions within $s(t)$ at a lower f). This can also be shown in a mathematically stringent way; the result is that the faster the variations in $s(t)$ the farther the spectrum extends into higher frequencies, and the more often $s(t)$ oscillates the more its spectrum concentrates at intermediate

*Equations (74-9) and (74-10) are similar to (74-4) and (74-5), but here based on the one-stage cutoff frequencies and not on corner frequencies.

†The higher the n the smaller the k; the variation of k with n is not much influenced by m. k is the square root of the quotient between the mn:th root of 2 minus 1 and the m:th root of 2 minus 1.

*Mathematically, the integral is the limiting case of a sum; when the period grows indefinitely, equation (74-6) approaches equation (74-13), in a certain sense. $s(t)$ is the integral from $-\infty$ to ∞, of s(t) times the exponential of j2πft, with respect to t.

†Regarding equation (74-13) as an integral equation, $S(f)$ can be solved into an unusually simple expression. Similar expressions hold for a_k and b_k in equation (74-6) (see literature).

‡Compare with equations (74-2) and (74-3).

§Compare with equation (74-2).

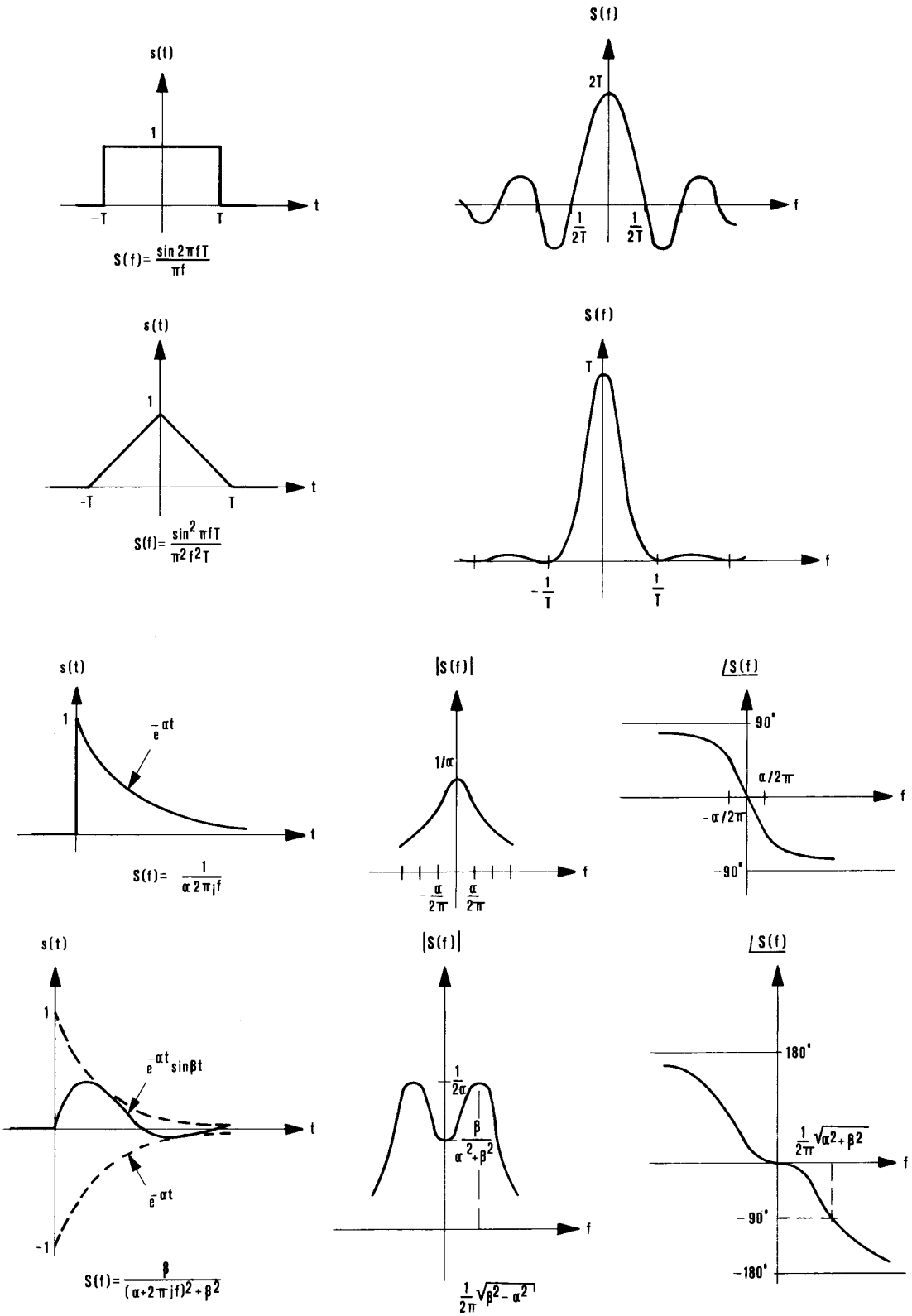


Fig. 74.6.—Some examples of spectra $S(f)$ of aperiodic time functions $s(t)$.

frequencies. The low-frequency extension of the spectrum is related to the total area under the signal $s(t)$; in the limit, putting $f = 0$ in equation (74-14) or (74-17):

$$S(O) = \int_{-\infty}^{\infty} f(t)dt \qquad (74\text{-}18)$$

i.e., the spectrum at zero frequency is related to the "d-c content" in the signal. One can show that, for low frequencies, the spectrum approximately follows

$$S(f) \approx \int_{-\infty}^{\infty} f(t)dt - j2\pi f \int_{-\infty}^{\infty} tf(t)dt \qquad (74\text{-}19)^*$$

in most cases. (An approximation for very high frequencies is much more difficult.)

As a highly oscillating signal has its spectrum largely concentrated around intermediate frequencies, a low-frequency cutoff well below these frequencies will not much influence the shape of the amplified signal, but the lack of d-c response causes a shift of the d-c level so that equation (74-18) is satisfied, i.e., the integral in equation (74-18) is made zero (of course no shift occurs if this integral is already zero for the input signal).

Some examples of spectra are shown in Figure 74-6.

Obviously, proper amplification of an aperiodic signal is achieved only if the amplifier lets the main part of $S(f)$ through without deforming it; that is, the main part of $S(f)$ should be contained in the passband of the amplifier. Below we will show this in terms of the time behavior of the am-

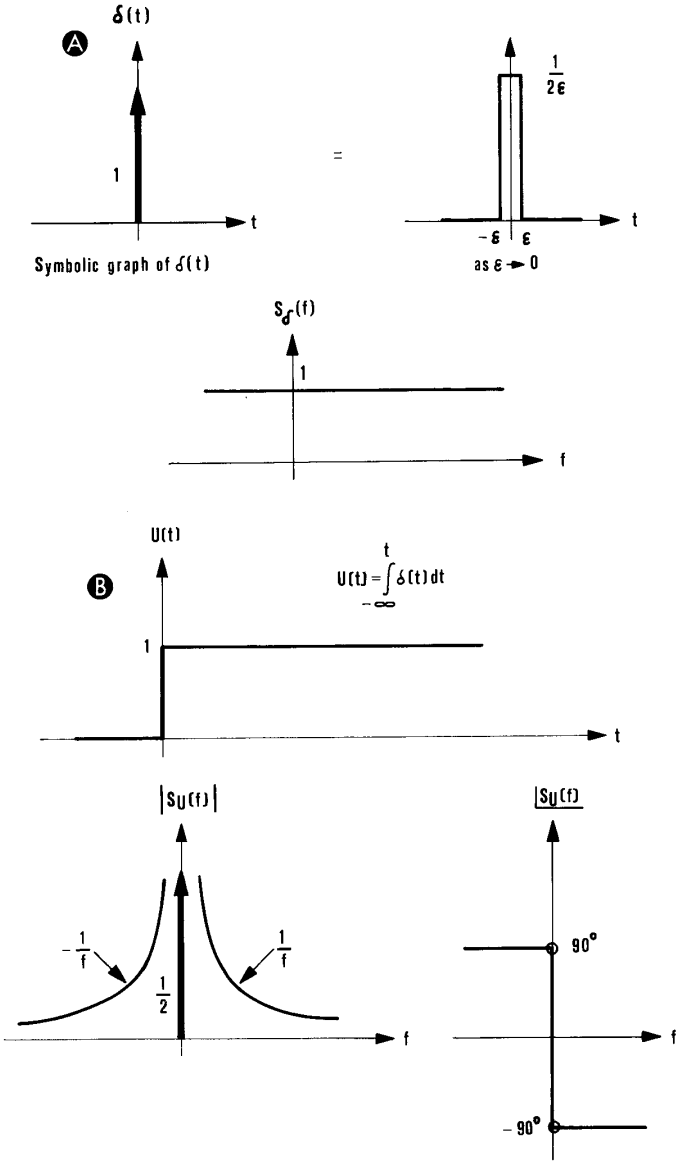


Fig. 74-7. – **A,** the impulse function and its spectrum. **B,** the step function and its spectrum.

*The two first terms in a power series expansion of $S(f)$ in equation (74-14).

plifier rather than its frequency behavior. For this purpose, we first study two special, fictitious time signals (as they never can be realized exactly in practice) of basic importance for the characterization of the time response of the amplifier; these signals are *impulses* and *steps*.

The Impulse Function and the Impulse Response

The impulse function $\delta(t)$ is a function "of zero duration and unity area," which can be thought of as the limit of the pulse sketched in Figure 74-7, *A*, as $\epsilon \to 0$. Its spectrum is constant:

$$\delta(t) \Rightarrow S_\delta(f) = 1 \qquad (74\text{-}20)^*$$

Hence, we see, from equation (74-16), that the transfer function $A(f)$ is the spectrum of the output signal when the input signal is an impulse function. The time function, corresponding to $A(f)$, is called the *impulse-function response*. If we denote this function by $a(t)$, one can show (applying the convolution theorem of the Fourier integral in equation [74-16]) that the output signal $s_o(t)$, corresponding to an input signal $s_i(t)$, is given by

$$s_o(t) = \int_{-\infty}^{\infty} s_i(x)\, a(t-x)dx = \int_{-\infty}^{\infty} s_i(t-x)\, a(x)\, dx \qquad (74\text{-}21)^\dagger$$

for any given input signal $s_i(t)$. Equation (74-21) (convolution of $f(t)$ and $a(t)$) expresses $s_o(t)$ as a *smoothed version* of $s_i(t)$ with the smoothing function $a(t)$; $s_o(t)$ comes out as a weighted mean value of $s_i(t)$ over the duration of $a(t)$, with $a(t)$ as weighting function. Therefore, it is important that $a(t)$ be short in duration as compared to important events in $s_i(t)$; otherwise, such events will be smoothed out and more or less erased. The shorter the duration of $a(t)$ the farther the extension of $A(f)$ on the frequency scale, as we have just found, looking at time-frequency relations.

Turning to low-frequency criteria, we find the following. If there is no d-c response in the amplifier, the impulse response will be (at least) biphasic or multiphasic or even a damped sinusoid, of the type sketched in Figure 74-8. From these sketches, we again see that a proper reproduction of low-frequency components in the input signal requires a sufficiently low low-frequency cutoff. If the low-frequency cutoff is too high up on the frequency scale, the negative portion of $a(t)$ tends to average out slow variations (as the integral over $a(t)$ is zero) if these occur during times longer than the duration of $a(t)$.

The Step Function and the Step Response

We first found that the impulse function response was related to the smoothing or weighted averaging of the signal. Here, we will show another response, related to the rounding off of fast transitions in the signal.

The step function $U(t)$ is a function of time that suddenly jumps from zero to unity at $t = 0$ but otherwise is constant. It is sketched in Figure 74-7, *B*, together with its spectrum $S_U(f)$, which is:

$$U(t) \Rightarrow S_U(f) = \frac{1}{2}\delta(f) + \frac{1}{j2\pi f} \qquad (74\text{-}22)$$

The step responses of the amplifier has, from equations (74-16) and (74-22), the spectrum

$$S_o(f) = \frac{1}{2}\delta(f)\, A(O) + \frac{A(f)}{j2\pi f} \qquad (74\text{-}23)$$

i.e., its spectrum is $A(f)$ weighted with $1/f$, except for constants and an impulse at the origin. One can show that this response, which we denote $a_U(t)$, is given by

$$a_U(t) = \begin{cases} \int_o^t a(t)\, dt, & t > 0 \\ 0, & t < 0 \end{cases} \qquad (74\text{-}24)^*$$

where $a(t)$ is the impulse response. The response of an amplifier to any input can be related to a_U. One can deduce, for the output signal $s_o(t)$ corresponding to an input signal $s_i(t)$ (its derivative is notated by $s_i'(t)$):

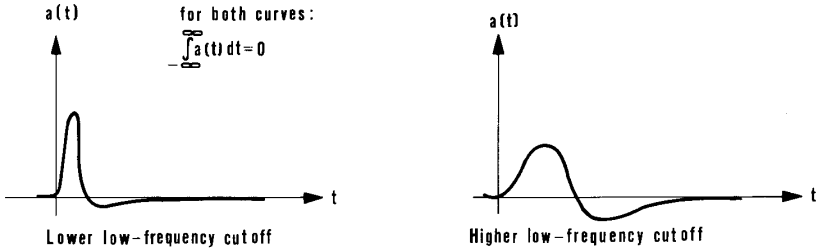
$$s_o(t) = \text{constant} + \int_{-\infty}^{\infty} s_i'(x)\, a_U(t-x)\, dx \qquad (74\text{-}25)^\dagger$$

where the constant is zero for an aperiodic signal applied at $t = 0$ (or any given time) and with a limited duration (or fading off to zero as $t \to \infty$). This means a smoothing of the derivative of $s_o(t)$, which may stand out clearer when equation (74-25) is differentiated (using equation [74-24]):

$$s_o'(t) = \int_{-\infty}^{\infty} s_i'(x)\, a(t-x)\, dx = \int_{-\infty}^{\infty} s_i'(t-x)\, a(x)\, dx \qquad (74\text{-}26)^\ddagger$$

corresponding to equation (74-21). A smoothing of the derivative results in a reduction and round-off of high derivatives and, hence, a slowing down of fast transitions in the signal. This will be discussed further in a somewhat different way below.

Fig. 74-8. – Examples of impulse function responses.



*$\delta(t)$ gives the spectrum $S_\delta(f) = 1$".

†One may understand this heuristically by regarding $s_i(t)$ as composed of a series of narrow pulses close to one another with corresponding amplitudes, each pulse adding its individual response to the output (integral is a limit of a sum).

*$U(t)$ is the integral of $\delta(t)$ and the same holds for the corresponding output signals. The integral may be taken from zero to t for t greater than zero; if t is less than zero, the $a_U(t)$ is zero anyway.

†As for equation (74-21), one can regard the input signal as composed of the sum of many small step functions, with amplitudes according to the derivative, each contributing an additive step response on the output.

‡Compare with equations (74-21) and (74-24).

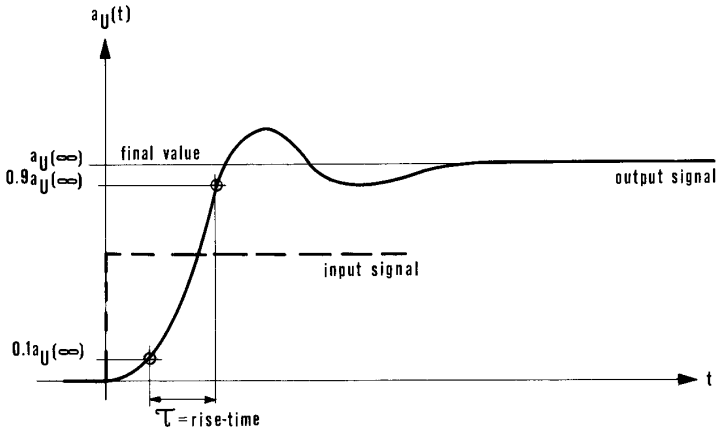


Fig. 74-9.—Example of a step function response with definition of rise-time τ for an amplifier with d-c response.

Rise-Time of the Amplifier

An important characteristic of the step response of an amplifier is the *rise-time*. This generally is defined, for an amplifier with d-c response, as the time it takes for the step response to rise from 10% to 90% of its final value (the limit as $t \rightarrow \infty$). This definition is illustrated in Figure 74-9. Some other definitions also exist (such as rise from 0 to 90% or other values, or the slope of the steepest rise; also a theoretically more appropriate but practically more difficult definition is related to the moments of the response) but will not be discussed here. The definition given above is, in practice, easy to handle and avoids some difficulties as to how the signal starts and how it approaches its final value.

If the amplifier has no d-c response, the final value is zero. In such cases, one extrapolates the curved large-time asymptote of the response to small times as a defined reference. As we are still discussing broad-band amplifiers, there is no problem, since the step response falls approximately linearly for intermediate times and, hence, a linear extrapolation will be sufficient. The rise-time then may be defined as the time required for a rise from 10% to 90% of the final value, measured "parallel" to the extrapolated linear asymptote, as sketched in Figure 74-10.

From the preceding discussion, we understand that the rise-time is related to the high-frequency behavior of the amplifier. Actually, one finds a relation between the rise time τ (in seconds) and the upper cutoff frequency f_u (in Hz):

$$\tau f_u = 0.30 \ldots 0.45 \qquad (74\text{-}27)^*$$

where the value of the product varies between the limits indicated, depending on the construction of the amplifier. The validity of this relation requires a small *overshoot*, less than 5%.

The overshoot is the amount by which the response at the end of the step exceeds its final or settled value—if it exceeds it at all; a system in which an overshoot exists is called undercritically damped; an overshoot-free system generally is overcritically damped. The limiting case, critical damping, is also free from overshoot. The overshoot is measured in percentage of the final value, as indicated in Figure 74-11. The existence of an overshoot and its amount as well as its form depend on the high-frequency behavior of the amplifier in the roll-off region. It may be oscillatory, as indicated in Figure 74-11 (appearing as a damped sinusoid superimposed on the response) or exponential, with one or a few "wiggles" (but a finite number of them), as in-

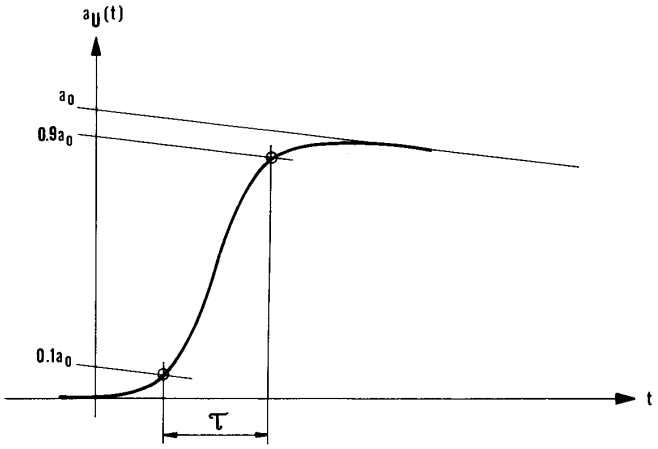


Fig. 74-10.—Another example of a step function response, showing a suitable definition of rise-time when the amplifier has no d-c response.

*This is empirically found to hold generally (when the overshoot is small) and has been mathematically proved only for special cases. The mathematical difficulties arise from a theoretically not very appropriate but practically easy definition of rise-time.

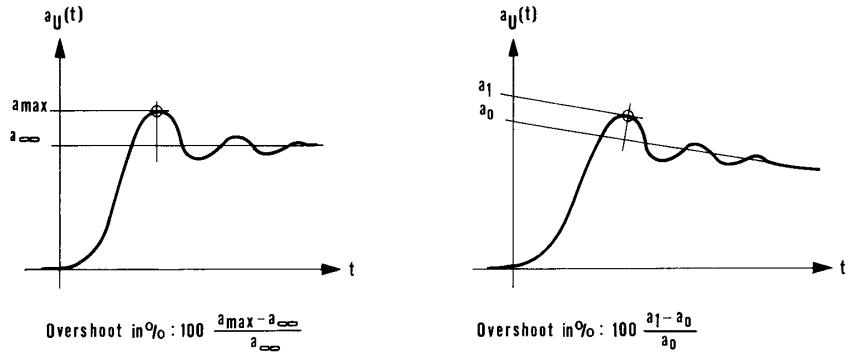


Fig. 74-11. – Further examples of step function responses, defining overshoot (to the left with d-c response, to the right without d-c response).

dicated in Figure 74-9. A *small* overshoot generally leads to a certain reduction of the rise-time without too much distortion of the response and, therefore, is desirable for some applications (e.g., in certain control systems and for pen-writer mechanisms).

Clearly, the rise-time is of vital importance when one needs amplification and wants to study fast transitions in the input signal. The amplifier can reproduce such transitions only if they occur at a much slower rate than the rise-time of the amplifier. If the input signal in the region of the transition is similar to a step response (directly, as in Fig. 74-9, or "obliquely," as in Fig. 74-10), i.e., remains fairly constant in a neighborhood on both sides of the transition, we may determine a rise-time of the input signal. One may approximately correct for the influence of the amplifier, evaluating the output signal by the rule of the combination of rise-times, which will be given below. This allows for a correction of the rate of the slope but not its form.

Combination of Rise-Times

As a general rule, the rise-times of n different amplifiers, $\tau_1, \tau_2, \ldots, \tau_n$, connected in *cascade*, combine to the total rise-time τ of the combined amplifier as

$$\tau \approx \sqrt{\tau_1^2 + \tau_2^2 + \ldots + \tau_n^2} \qquad (74\text{-}28)^*$$

if the overshoot is less than 2%. This approximation is better as n becomes higher. As mentioned above, the rise-time of the input signal (if this signal is such that a rise-time can be defined for it), combines in the same way:

$$\tau_o \approx \sqrt{\tau^2 + \tau_i^2} \qquad (74\text{-}29)^\dagger$$

if τ_i and τ_o are the rise-times of the input and output signals and τ that of the amplifier. Obviously, a rise-time correction according to equation (74-29) can be performed only as long as τ_i is not much smaller than τ.

The rise-time is a measure of the smoothing effect on sharp transitions, mentioned earlier, and the effect is expressed by equation (74-29). The overshoot is a measure of ringing in the amplifier – if present – which also was discussed earlier.

The Reproduction of a Pulse

Due to the high-frequency behavior, a square pulse at the input will be affected by the rise-time and, possibly, by overshoot of the amplifier; these effects will be seen at the output. Due to the low-frequency behavior, the output pulse will not have a horizontal "roof" or "top" but be falling gradually, after the rise, unless the amplifier has d-c response. These effects lead to a distortion of the pulse as sketched in Figure 74-12 for different cases. Several descriptive parameters other than rise-time and overshoot may be defined. The *delay* of the pulse is defined in the first curve in Figure 74-12. In obvious analogy, one may also define fall-time and undershoot at the falling part of the pulse response. The fall of the "roof" of the pulse, the *pulse fall*, is best characterized by its *time constant*, defined in Figure 74-13, as given by the inital derivative of the falling curve.

The Influence of the Phase Characteristic

This influence is illustrated in Figure 74-5 for a periodic signal and can lead to a similar distortion for an aperiodic signal. The phase-amplitude relations at high-frequency cutoff and roll-off can be of vital importance for the existence and amount of an overshoot. Here, only one special case will be discussed, that of the linear phase characteristic. If the transfer function (amplification) has a phase, varying linearly with frequency, i.e.,

$$A(f) = |A|e^{-jkf} \qquad (74\text{-}30)$$

where k is a constant within the passband (although it cannot be realized for all frequencies), the output signal for an input signal with its spectrum well within the passband can be written (from equations [74-13] and [74-16]):

$$s_o(t) \approx |A| \int_{-\infty}^{\infty} S_i(f) e^{j2\pi\left(t-\frac{k}{2\pi}\right)f}\, df \qquad (74\text{-}31)^*$$

or, from equation (74-13):

$$s_o(t) \approx |A|\, s_i\left(t - \frac{k}{2\pi}\right) \qquad (74\text{-}32)$$

i.e., the output signal is merely a *time-shifted* reproduction of the input signal (except for the constant factor $|A|$). We therefore see that a linear phase characteristic does not cause a distortion but only a delay.

General Discussion

It can be seen from the above that special care should be taken when judging the rise and fall, the delay and the over-

*This arises from an analogy with statistics and is analogous to the central limit theorem. Equation (74-28) holds exactly as $n \to \infty$.

†Consider the signal as the output from a fictitious amplifier with rise-time τ_i and a step on the input.

*k in equation (74-30) is constant only within the passband, but if $S_i(f) \neq 0$ essentially only in the passband, equation (74-31) holds to a good approximation.

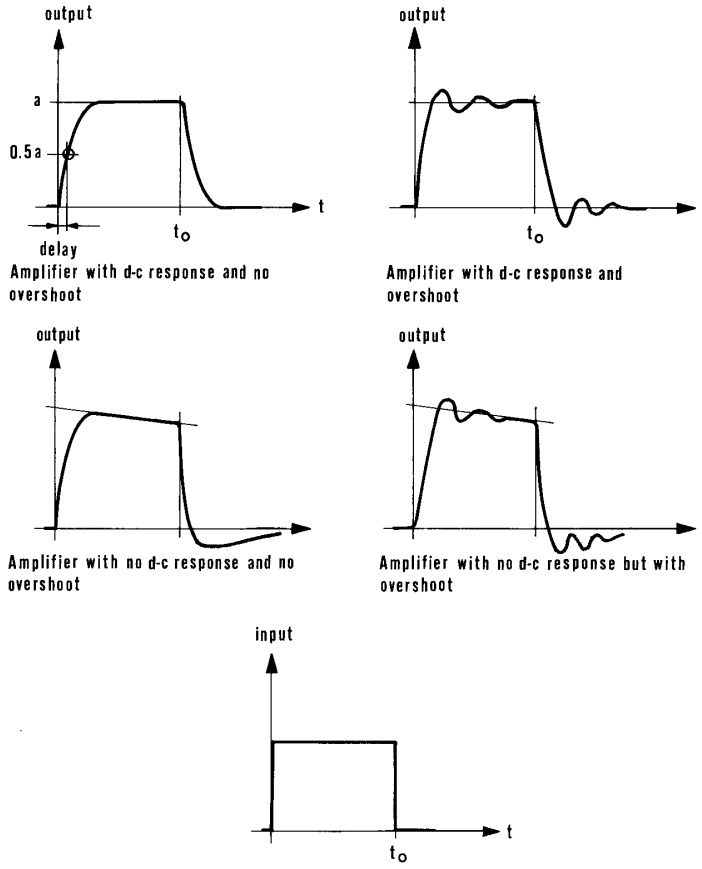
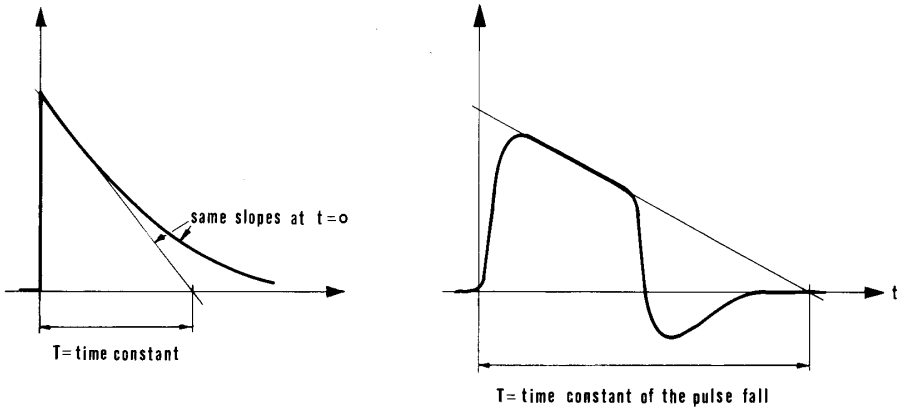


Fig. 74-12. – Reproduction of a square pulse in different cases, including definition of delay.

all resultant form of aperiodic signals that have been passed through amplifiers. These appearances may, under certain circumstances, be largely influenced and sometimes falsified by the amplifier itself; this can lead to an underestimation of the steepness of a transition in the signal or an overestimation of a delay as well as a misjudgment of the form. Furthermore, "wiggles" may be caused by the amplifier and certain details (short-time events or details of fast transitions) may be erased by the smoothing effect of the nonideal impulse response. The proper representation of the input is arrived at only when the amplifier has a sufficiently high upper cutoff frequency, as related to the spectrum of the signal, and d-c response. For quickly varying or short signals, the d-c or low-frequency response may be of minor

Fig. 74-13. – Definition of a time constant in general and, specifically, for the pulse fall.



importance as long as the corresponding lower cutoff frequency is considerably below the main spectrum of the signal. For a highly oscillatory signal, the lack of d-c response merely shifts the d-c level.

The phase response may also be a source of signal distortion, unless the phase remains reasonably constant within the passband. As was seen, a linearly varying phase may also be accepted, as this does not influence the form but only causes a measurable and consistent delay. The realization of an approximately linear phase characteristic within the passband affects the roll-offs outside—generally far outside—the passband. This holds especially for so-called minimum-phase transfer functions. Non-minimum-phase transfer functions may be thought of as minimum-phase functions combined with all-pass filters; such all-pass filters may aid in linearization of the phase characteristic.

Bandpass Amplifiers and Filters

Above, we discussed the properties of and requirements for an amplifier that should reproduce the input signal as truly as possible within practical tolerance limits. Here, we will discuss briefly the properties of amplifiers with characteristics tailored to extract certain features from the signal, such as a part of the spectrum or a certain sinusoidal component, or to process the signal, e.g., to obtain its derivative or integral.

Bandpass Amplifiers

These are amplifiers used to extract a certain part of the spectrum of available signals. If relatively broad-band in nature, they doubtlessly have, in general, the same properties as broad-band amplifiers. The main difference is in their use; they are not intended to let the whole spectrum through but only a defined part of it, *intentionally* causing an output signal to appear as if the input signal had been subject to the distortion discussed earlier in this chapter. With this intention, one usually desires as sharp a cutoff as possible at the ends of the passband; this usually calls for a linear phase characteristic in the passband, as a constant characteristic is quite difficult to realize. The most important use of such amplifiers is where the input signal has two or more components, one of which has the main part of its spectrum in a region where the other signals contribute comparatively little. A bandpass amplifier, with its passband in this region only, then can separate the first signal from the others; i.e., it acts as a *filter*.

A special type of bandpass amplifier is the narrow-band amplifier, discussed below, which has such a narrow bandwidth and sharp cutoffs that it almost filters out a single sinusoidal component in the available spectrum.

Other special cases are high-pass and low-pass amplifiers and bandstop amplifiers, also discussed below.

Narrow-Band Amplifier

The narrow-band amplifier or selective amplifier is designed to amplify signals only within a surrounding of a certain frequency, its *center frequency*. It may be used to indicate the mere existence of a sinusoidal component and should then, in principle, have as small a bandwidth as possible. It may also be used to filter out a time-varying sinusoidal component and then must have a bandwidth that is sufficiently broad to reproduce the time variations truly at its output. The latter requirement will be discussed briefly.

A sinusoidal component may have a time-varying amplitude or time-varying frequency or both. In the former, an *amplitude-modulated signal*, the basis for the bandwidth requirement is as follows. A sinusoidal signal $\cos(2\pi f_0 t + \varphi_0)$, modulated with a time-varying amplitude $m(t)$,

$$s(t) = m(t)\cos(2\pi f_0 t + \varphi_0) \qquad (74\text{-}33)$$

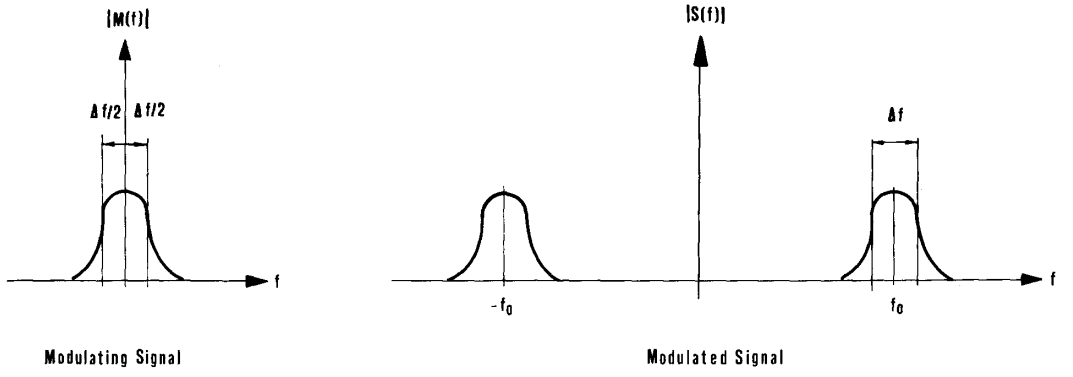
has the spectrum

$$S(f) = \frac{1}{2}[M(f - f_0)e^{j\varphi_0} + M(f + f_0)e^{-j\varphi_0}] \qquad (74\text{-}34)^*$$

i.e., the spectrum of the modulated signal is composed of the displacements of the spectrum $M(f)$ of the modulating signal $m(t)$ to the frequencies f_0 and $-f_0$, as sketched in Figure 74-14. Normally, the spectrum of $m(t)$ extends over frequencies much less than f_0. (Note the two-sided spectrum of $M(f)$; equation [74-14] also gives a spectrum for negative frequencies, which have no direct physical meaning, so that $|M(f)|$ is an even function, but here also the negative frequencies in $M(f)$ contribute to $S(f)$ through shifting of $M(f)$.) Thus, it is important that the narrow-band amplifier has sufficiently constant amplification over a frequency range Δf around f_0, as broad as the main part of the spectrum of the modulating signal (see Fig. 74-14). Within this range, it should also have a linear phase characteristic.

If the signal is frequency modulated, the situation is con-

Fig. 74-14.—Spectra for an amplitude-modulated sinusoid, $S(f)$, and its modulating signal, $M(f)$.



*$M(t)$ is shifted not only in frequency but also in phase, as indicated by the exponential factors.

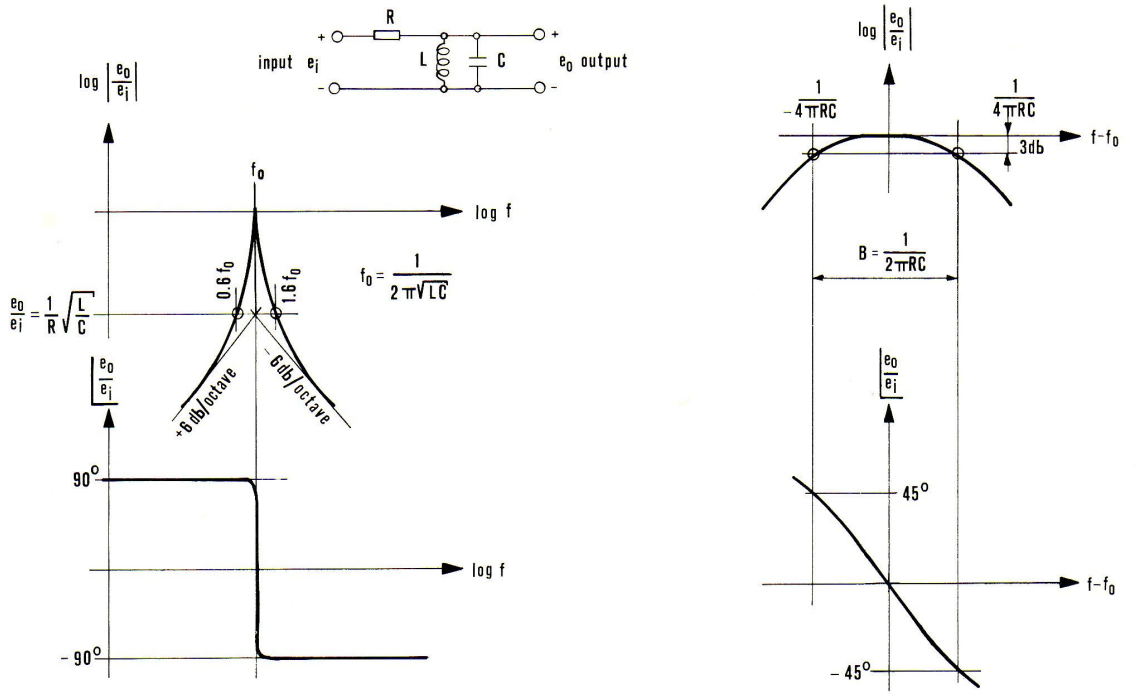


Fig. 74-15. – Characteristic of a resonant circuit.

siderably more complicated. Even in the simple case of a sinusoidal modulating signal, the modulated signal is expressed by an infinite number of sinusoidal components, distributed around f_o and $-f_o$ and spaced apart by the frequency of the modulating signal (the amplitudes of these components are given by Bessel functions). It turns out that for a frequency-modulated signal

$$s(t) = m_o \sin \{2\pi[f_o + g(t)]t + \varphi_o\} \qquad (74\text{-}35)^\dagger$$

it generally is sufficient to have a bandwidth of

$$B \approx 2(g_d + \hat{f}) \qquad (74\text{-}36)$$

where the "frequency deviation" g_d is the maximal value of $|g(t)|$ and $\hat{f}$ the upper frequency limit for the important part of the *spectrum of* $g(t)$. The narrow-band amplifier stage often is built around a resonant circuit similar to the one sketched in Figure 74-15, with its transfer characteristic. The amplifier then has approximately the same characteristic. Note the acceptable linearity of the phase characteristic. Modern circuitry also often makes use of "active filters" realizing similar characteristics through feedback (or one uses ceramic resonant components or rapid switching between different circuit paths). Often different stages are used in cascade to reduce the bandwidth – or to obtain a more flat passband together with steeper roll-offs by centering the individual stages at slightly different center frequencies.

High-Pass Amplifier

For some purposes, it is desirable to remove low-frequency components in the signal. For such purposes, one uses a broad-band amplifier with its low-frequency cutoff above the uninteresting frequency region and a steep roll-off. Nature puts high-frequency cutoff as well, anyway, but often it is desired to place this cutoff intentionally at a certain frequency. Here, the discussion of broad-band amplifiers holds, for the undesired signal, especially as to operation on the low-frequency roll-off.

A special case is the differentiator, which has a linearly sloping (rising) low-frequency characteristic (roll-off), i.e., an amplification proportinal to f, up to a certain frequency where nature or intention puts a high-frequency cutoff, usually without a passband in between. It can be shown that the output signal is the derivative of the input signal by differentiating equation (74-13) with respect to time (left as an easy exercise for the reader), as long as the important part of the spectrum of the input signal falls on the linear slope.

Low-Pass Amplifier

If, instead, it is desired to leave out an undesired high-frequency component, one uses an amplifier with a d-c response, an upper cutoff frequency that falls below the uninteresting region and a steep roll-off. These amplifiers behave as discussed above for broad-band amplifiers with d-c response, for the undesired signal especially, as to operation in the high-frequency roll-off region.

A special case is the integrator, which has a linearly falling roll-off slope, i.e., an amplification proportional to $1/f$, at least up to a determined frequency (usually the slope is increased after a certain frequency). It can be shown by integrating equation (74-13) with respect to time (also left to the reader as an easy exercise), that the output signal is the integral of the input signal (as long as the slope does not change over the main part of the spectrum).

†Equation (74-35) is the general expression for a frequency-modulated signal, i.e., having a time-varying frequency, but can also be shown to be equivalent to a phase-modulated signal (time-varying phase), with a proper $g(t)$.

General Discussion

Bandpass amplifiers with a relatively broad passband do not differ considerably from broad-band amplifiers. The difference is mainly in their use (and that roll-offs usually are made steeper).

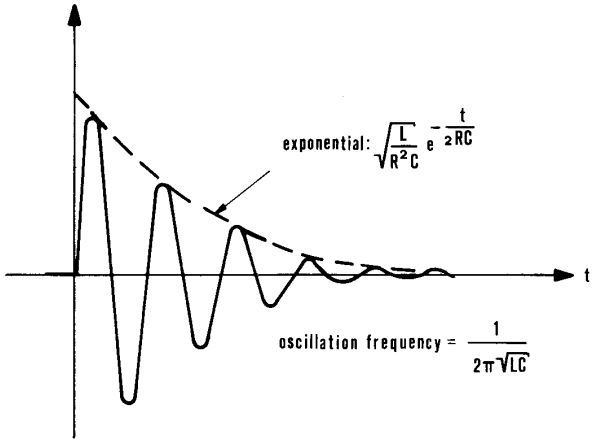
The special case of narrow-band amplifiers, used for studying single sinusoidal components, shows that the bandwidth must be tailored to the amplitude or frequency variations one wishes to study. Too low a bandwidth has an effect on amplitude modulation analogous to that of a broad-band amplifier with d-c response, but too low a bandwidth, applied on the modulating signal alone. The effect of too low a bandwidth on a frequency-varying signal is considerably more complicated.

It should be mentioned that a narrow-band amplifier, of course, also reacts to the portion of the spectrum of an aperiodic signal, which falls in the passband and on the nearby roll-offs. As an example, the step-function response of the LC filter in Figure 74-15 is shown in Figure 74-16. This could have been taken for an amplitude-modulated sinusoid with an exponential modulation (of short duration). Hence, it should be kept in mind that transient signals (and noise!) can appear much like modulated sinusoids on the output of a narrow-band amplifier and, under certain circumstances, be mistaken for such. It seems likely that this may occur when EEG signals are interpreted (often containing spikes or fast transitions), e.g., alpha waves of very short duration may be "indicated" at the filter output but actually be caused by irregular transitions or spikes.

Finally, when using a differentiator or integrator, one must, for obvious reasons, be sure that the important part of the spectrum is kept on the proper linear slope of the transfer function. This is especially important for differentiators, almost abruptly changing from a rising to a falling slope at a certain frequency; deviation in the phase characteristic from the ideal, linearly sloping phase may be relatively large even at a tenth of this frequency.

In general, high-pass and low-pass amplifiers can, under certain circumstances, deliver distorted versions of the undesired signal component (due to the presence of this component at the roll-off), which can interfere with the desired component. In any case, it is important to make sure that the desired signal has the important part of its spectrum within the passband.

Fig. 74-16. – Step-function response of the circuit in Figure 74-15.



Bandstop Amplifiers

A bandstop (band rejection or "notch") amplifier has specifically low amplification (at or near zero) in a certain frequency range and passbands on one or both sides of this range, with roll-offs between. With a broad bandstop range, this may be looked at as a broad-band amplifier with one or two passbands. Often a narrow bandstop range is used, e.g., at 50 or 60 Hz to remove mains interference. This has little influence on the amplitude spectrum of signals with broad spectra, but may have an important influence on the phase spectrum. The range of phase variation in the transfer function is considerably broader than the bandstop range.

Noise

Here, we will discuss noise in its stricter and more appropriate sense – randomly varying signals. Although interfering and undesirable signals of non-random nature sometimes are called "noise," the proper term should be *interference*.

Noise is a random fluctuation, meaning that the momentary value of, for example, a noise voltage, cannot be predicted exactly but that one knows the probability that it falls between given limits. Hence, noise is *unpredictable*, in contrast to other interfering signals, such as power line "hum," which are *deterministic* – the momentary values of which are determined by a certain law.

Thermal Noise

Noise arises, in principle, in all conductive and current-carrying media in electric circuits. Noise originating from connecting wires is negligible, as is also usually the case for noise in capacitors and inductors. The noise stemming from resistors, on the other hand, is of basic importance. A noise voltage appears across the connections of the resistor, due to thermal fluctuations in the electron distribution within the resistive material. This is called *thermal noise*.

When trying to find the spectrum of a noise, one runs into certain mathematical difficulties due to the random nature and constant power of the signal. With a certain modification of the definition of the spectrum (see above), one can define a *power spectrum* for the noise, analogous to $|S(f)|^2$ for the spectrum $S(f)$ of a deterministic signal $s(t)$ (applied on a deterministic signal, the power spectrum as defined for noise is essentially the same as $|S(f)|^2$). However, a phase spectrum cannot be defined. For details concerning noise power spectrum, the reader is referred to the literature.

The power spectrum of thermal noise of a resistor of R ohms is 4 kTR (volts)^2 per Hz (when defined for positive frequencies only), where k is Boltzmann's constant ($1.38 \cdot 10^{-23}$ $Ws/°K$) and T is the absolute temperature in °K. If the rms voltage E of this noise is measured, connecting a noise source and a measuring device to an amplifier with a bandwidth of B Hz, one finds

$$E = 2\sqrt{kTRB} \qquad (74\text{-}37)^*$$

if B is so large or the roll-offs so steep that the influence of the roll-offs can be ignored. Hence, the effective part of the input rms thermal noise grows proportionally to the square root of the bandwidth of the system through which the signal is processed.

Noise often is stated in terms of its "peak-to-peak" value. This is an inappropriate term, as theory shows that a

*Note the dependence on temperature. This is why sometimes certain electronic equipment is cooled by liquefied gases where the requirements are extreme, such as in radio-astronomy.

true peak value for noise does not exist. In discussing "peak-to-peak" value, we mean the range within which the momentary value falls 97.3% of the time (triple standard deviation range). It can be shown that this is related to the rms value as

$$(\text{rms value}) = \frac{1}{6}(\text{"peak-to-peak" value of the noise}) \qquad (74\text{-}38)\dagger$$

Measuring "peak-to-peak" values on an oscilloscope, one finds approximately six times the rms value, as the probability for the signal to exceed this range is quite small and therefore higher momentary variations are almost never seen on the oscilloscope.

One should note that the noise over a general physical impedance has the rms value

$$E = 2\sqrt{kT\int_0^\infty R(f)\,df} \qquad (74\text{-}39)\ddagger$$

where $R(f)$ is the real part of the impedance. For a resistance R and a capacitance C in parallel, for example, one finds

$$E = \sqrt{\frac{kT}{C}} \qquad (74\text{-}40)$$

This seems to indicate that the capacitor is the source of the noise (as R does not appear in the formula). This is not true but results from the filtering action of C on the physical noise source, which is the resistor. Equation (74-39) gives the general result of the filtering action of the circuit itself on its resistive noise sources, as any physical impedance is associated with a network of limited bandwidth. (In equation [74-37], on the other hand, the theoretic case of a pure resistance as the source was considered, which would give an infinite rms value—as pure resistance is associated with a circuit of infinite bandwidth—and therefore a systematic bandwidth had to be introduced.)

Because of the uniform power spectrum of thermal noise, it often is called *white noise* (in analogy with white light).

Noise in Semiconductors

In an active circuit, one also has noise contributions from transistors and often also from semiconductor diodes. The theory of noise in vacuum tubes and diodes will not be explained, since these rarely are used today. Generally, one finds a noise contribution that is caused by random fluctuations in the "d-c" current through the device. For a semiconductor diode in forward conduction operation, one finds the rms value for the noise current:

$$I = \sqrt{2q(2I_o + I_{d\text{-}c})B} \qquad (74\text{-}41)$$

where q is the electron charge ($1.6 \cdot 10^{-19}$ As), $I_{d\text{-}c}$ the d-c current through the diode in amperes, I_o the limit leakage current in amperes in backward operation (for high backward voltages not causing avalanche breakdown) and B the bandwidth in Hz of the measuring instrument, giving I in amperes. Again, this is a white noise. To this is added *flicker noise* with a noise spectrum proportional to $1/f$, i.e., inversely proportional to frequency. Because of this form of the spectrum, it is also often called *1/f noise*. This noise contribution is mainly caused by phenomena occurring on the surface of the semiconductor material and phenomena associated with the recombination of charge carriers ("holes" and electrons). The spectrum for the total noise current in a diode is sketched in Figure 74-17 for low and medium frequencies (at higher frequencies, additional effects influence the spectrum, such as the influence of layer and parasitic capacitances or lead inductance).

The noise in a transistor is something much more complicated, and any detailed discussion falls beyond the scope of this chapter. For low and intermediate frequencies, the spectrum of the collector noise current has the same appearance as that of a diode (see Fig. 74-17). The characteristic at higher frequencies will be mentioned later in conjunction with the noise factor.

Noise Factor and Signal-to-Noise Ratio

Amplifiers and transistors have a general specification in terms of a noise factor. Several special types of noise factors and noise measures have been defined for different purposes; only one concept will be mentioned here.

Suppose that the amplifier input impedance and the source impedance are matched, meaning maximal transfer of power from source to amplifier (as will be discussed in

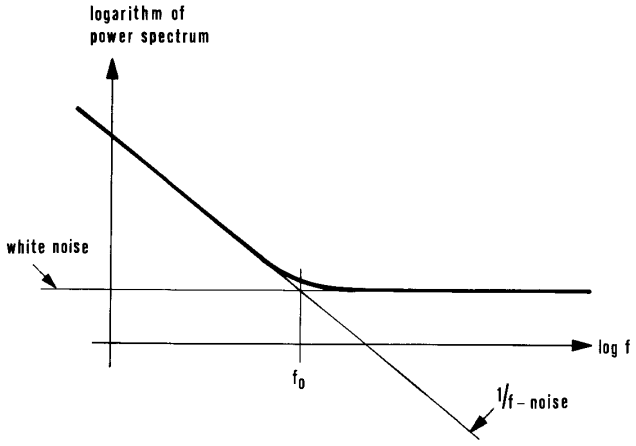


Fig. 74-17.—Power spectrum of the noise current in a semiconductor diode.

†Note that this holds for only this type of noise and is not a general relation.

‡Equation (74-39) obviously is a generalization of equation (74-37) in the case when "B" is determined by the impedance itself and not by the "external" circuitry.

the section on input and output characteristics) and that the load on the amplifier at its output is in the same way matched to its output impedance, we connect a resistor as a noise source to the input and measure the resulting power spectra in the load. For a real amplifier, one measures a total power spectrum as a function of frequency, $W_t(f)$, in the load, which is contributed to by the noise source at the input *and* noise sources inside the amplifier. If the amplifier were ideal (i.e., without any internal noise source), we would measure a reduced power spectrum $W(f)$ in the load. The *noise factor* or *noise figure* then is defined as

$$F(f) = \frac{W_t(f)}{W(f)} \qquad (74\text{-}42)$$

characterizing the "noisiness" of the amplifier, as referred to a "standard" resistive source, in the case of maximal power transfer from source at the input to load at the output. (More exactly, F as defined above is called the "standard operating noise figure" to avoid confusion with other types of noise figures, defined for special purposes.)

$W_t(f)$ may be divided into two additive parts, the contribution $W(f)$ from the input source alone and the contribution $W_i(f)$ from the internal sources in the amplifier. One then can write

$$F(f) = 1 + \frac{W_i(f)}{W(f)} \qquad (74\text{-}43)^*$$

which will be used later when discussing cascaded stages. We see that $F = 1$ for an ideal amplifier and $F > 1$ for a real amplifier.

If $W_s(f)$ is the power delivered to the amplifier from the source, we may define a power amplification in full analogy with the definition of amplification given earlier, choosing power instead of voltage as an input and output quantity. If the power amplification is denoted $G(f)$, we then have $W(f) = W_s(f)\ G(f)$. As we have assumed matching impedances for the highest possible transfer of power, $G(f)$ is called *available power* gain. This concept is necessary for later discussion.

The *signal-to-noise ratio* (SNR or S/N) is a useful measure of the "noisiness" of a signal, defined as the quotient of the rms value of the signal alone (the noise-free signal) and that of the noise part of the composed signal. The signal-to-noise ratio of an amplifier is the signal-to-noise ratio of its output signal for a given input signal; for a bandpass amplifier the ratio often is expressed in db assuming, for example, an input signal of 1 μV rms value. The SNR of broadband amplifiers depends on the frequency range if the passband covers regions with $1/f$ noise or with a non-constant high-frequency noise characteristic. Another, frequently used method of characterizing noise properties of an amplifier is to *refer its inherent noise to the input.* One then states the rms value of an input noise source that gives the same relative noise contribution, added to the input signal, as appears in fact at the output. This rms noise applied to the input of the idealized amplifier – the identical but noise-free amplifier – gives the same noise at its output as is measured at the output of the real amplifier with no input signal.

Noise Factor of a Transistor

A typical frequency dependence of the noise factor of a transistor is shown in Figure 74-18, where the effect of $1/f$ noise and the behavior at high frequencies (mainly due to reduction of the available power gain at high frequencies) appear. Note that F is a relation between power spectra, the definition of the *db unit for power relations* gives the expression for F in db stated in Figure 74-18 (compare with the first part of this chapter, where db for voltage or current relations is defined). Hence, a slope proportional to f or $1/f$ corresponds to a slope of ± 3 db/octave in "power decibels" and 6 db/octave in "voltage decibels."

An amplifier generally will have a noise characteristic of the same type.

Chopper Amplifier

The $1/f$ noise exists in any normal amplifier with d-c response. To avoid the influence of this noise, one can shift a "d-c" signal (pure d-c or slowly varying) in frequency by modulating an a-c signal with it (compare equations [74-33] and [74-34]). The modulated a-c signal then is amplified in an a-c amplifier, not contributing $1/f$ noise; the amplified (modulating), desired signal then is extracted with a detector. The simplest way to do this is to "chop" the signal by multiplying it with a square wave, having a frequency that is much higher than the relevant spectral range of the desired signal. This square wave varies between either $+1$ and -1 (periodic sign inversions of the desired signal) or $+1$ and 0 ("sampling") and the input to the a-c amplifier, therefore, is an amplitude-modulated square wave. After amplification, it can be detected (the desired signal is extracted) by multiplying it again with a square wave and filtering to remove the remaining high-frequency a-c. Of

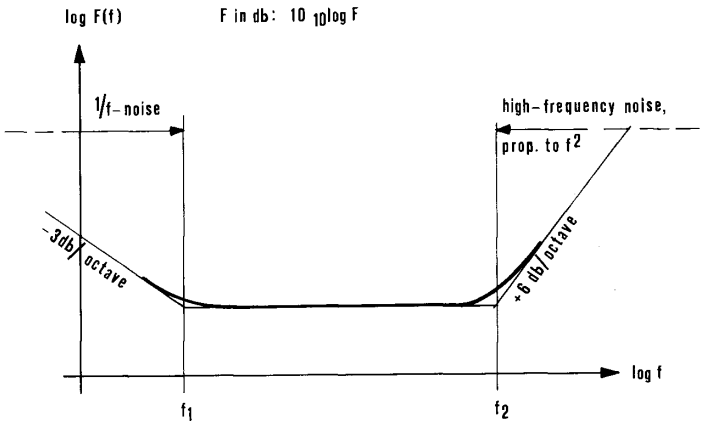


Fig. 74-18. – Noise factor of a transistor as a function of frequency.

*As, by definition, $W_t = W + W_i$.

course, the "chopping" must be designed so that it does not contribute serious *l/f* noise itself.

This technique is applied in *chopper amplifiers* for low-noise d-c amplification. A *parametric amplifier* with d-c response works with a somewhat related principle (using voltage-controlled capacitors).

Cascaded Amplifiers. The Importance of the First Stage

If two amplifiers are connected in cascade, the total noise factor, with a "standard" resistive noise source at the input, is

$$F = 1 + \frac{W_i}{W} = 1 + \frac{W_{i2} + G_2 W_{i1}}{G_1 G_2 W_s} \qquad (74\text{-}44)$$

where W_{i1} and W_{i2} come from the internal noise sources and G_1 and G_2 are the available power gains of the first and the second amplifiers, respectively. Obviously, the part W_{i1} of the power from the first amplifier contributes by $G_2 W_{i1}$ to the output of the second amplifier. The over-all power gain is $G_1 G_2$, as is easily realized.

Defining the individual noise factors F_1 and F_2 of the amplifiers according to equation (74-43), we find

$$F = F_1 + \frac{1}{G_1}(F_2 - 1) \qquad (74\text{-}45)$$

By applying this formula in sequence on a cascade of n amplifiers, one finds

$$F = F_1 + \frac{F_2 - 1}{G_1} + \frac{F_3 - 1}{G_1 G_2} + \cdots + \frac{F_n - 1}{G_1 G_2 \cdots G_{n-1}} \qquad (74\text{-}46)$$

The available power gain for each amplifier generally is greater than unity. We see, therefore, that to keep F as small as possible, it is important to have (1) the lowest possible noise factor in the *first* stage and (2) the highest possible gain in the *first* stage. Obviously, the first stage generally determines the noise properties of the whole cascade. Hence, efforts to provide a low-noise amplifier should concentrate on this stage.

Detection of Noisy Signals

Amplitude detection of a periodic signal generally is performed with a diode circuit (see Chapter 71). As a diode characteristic is not exactly piecewise linear but rounded off around zero, the detection of small signals with noise can result in a reduction of the signal-to-noise ratio. If the diode characteristic is approximated by a parabola in its curved region, one can show that for signals detected in this region the signal-to-noise ratios behave as:

$$\begin{cases} \left(\frac{S}{N}\right)_{out} \propto \left(\frac{S}{N}\right)_{in}, \text{ if } \left(\frac{S}{N}\right) \gg 1, \\ \left(\frac{S}{N}\right)_{out} \propto \left(\frac{S}{N}\right)^2_{in}, \text{ if } \left(\frac{S}{N}\right) \ll 1 \end{cases} \qquad (74\text{-}47)$$

Hence, a bad signal-to-noise ratio becomes even worse after the detection. It is, therefore, important that such signals have sufficient amplification before detection, so that the detection can be made on a sufficiently linear part of the diode characteristic.

Bandwidth and Signal-to-Noise Ratio

We have seen that the spectrum of the noise, contributed by the amplifier, is constant within a broad region and rises at low frequencies. When amplifying small signals, it is therefore important that the amplifier does not have too large a bandwidth, as the parts of the passband outside the range of the important part of the signal spectrum contribute to noise but not to the signal. Generally, the filtering effect of the amplifier on its own noise is quite similar to that on the input signal, as it is the first stage that determines the noise characteristic and the later stages often determine the bandwidth. If an important band-limiting occurs at the input to the first stage (as often is the case for high-input impedance amplifiers), this band-limiting should be repeated in a later stage to act on the noise of the first stage as well as on the signal.

Assuming, for simplicity, purely white noise, the rms noise voltage on the output is proportional to $\sqrt{B}$, where B is the bandwidth, but the signal rms voltage remains constant when B increases over a certain value; before this value is reached, it also increases. Therefore, an optimal bandwidth can be found, giving the highest signal-to-noise ratio. For noisy signals, it can be important to make the bandwidth optimal. This generally is near the width of the important part of the signal spectrum, but if it falls off slowly, the optimal signal-to-noise ratio may require a certain distortion of the signal spectrum. In such cases, the proper reproduction of the input signal at the output also must be included as a proper factor in the optimization.

Drift in D-C Amplifiers

Drift in d-c amplifiers often can be considered as a special case of *l/f* noise. Drift has been found to follow the *l/f* law down to corresponding frequencies in the "drift spectrum" as low as $6 \cdot 10^{-5}$ Hz, corresponding to a period of 5 hours, and one might expect this to hold for considerably longer drift periods. The use of "d-c" feedback to stabilize against drift can, in this context, be looked on as providing a cutoff at very low frequencies.

Narrow-Band Amplifiers

Of the continuous noise spectrum, spread over all frequencies, a narrow-band amplifier filters out a narrow portion, corresponding to its passband. This may appear as an amplitude *and* frequency-modulated sinusoid at the output (see equation [74-39]). When studying, for example, the α component of the EEG by filtering, one has a contribution from the EEG itself (with its noisy appearance) even if an α wave is not present. If amplitude *and* frequency (within the passband) both seem to vary statistically with time within shorter time periods, one might suspect that one observes filtered "EEG noise" rather than an α wave. The best test probably is to use several filters at different center frequencies in parallel, i.e., to observe other EEG waves simultaneously even if one is interested only in the α wave. If all filters have similar signals (except for different center frequencies) at their outputs, one can expect that this is contributed to by the noisy appearance of the EEG alone and does not correspond to sinusoidal components in the EEG. Only when one or two filters have a markedly different output signal, as compared to the others, can one expect the EEG to have a corresponding sinusoidal component. This holds also for the influence of spikes and fast transition in the EEG, mentioned earlier. It seems especially useful if one would have a "test filter" centered on a frequency where rarely sinusoidal components appear in the EEG, in order to compare its output with other filter outputs.

Averaging

Aperiodic signals can be extracted out of noise through *averaging*, if they are repeated many times. To do this, one,

in principle, adds all the aperiodic signals. As the mean value of the noise is zero, this means that the noise contribution tends to average out, whereas the repeated aperiodic signal is built up proportional to the number of repetitions. If n repetitions of the signal are added, the sum signal has a signal-to-noise ratio improved by a factor of $\sqrt{n}$.

Other Noise Sources

A special type of transistor noise, which has become actual with the development of integrated circuits, is the *burst noise*. This appears almost as a square wave with statistically varying frequency, i.e., the noise voltage (or current) switches almost between two levels.

A noise contribution from a resistor, which often is forgotten, is its *current noise*. In addition to the thermal noise, mentioned above, a current noise appears when a d-c current flows through the resistor, which is due to statistical fluctuations in the current distribution in the resistor. This noise is of the $1/f$ type and the appearing noise voltage has a rms value proportional to the d-c current. It is, furthermore, strongly dependent on the type of resistor used; generally, metal film resistors are the best. Hence, a low-noise amplifier should employ high-quality resistors (for these the manufacturer usually specifies the current noise, generally in μV/V, meaning μV rms noise per volt d-c over the resistor, measured over a specified frequency range) in its first stage and also run low d-c currents through them.

Stability

This part presents a brief discussion of the stability of an amplifier in the sense that it should not generate a signal of its own; the output of an amplifier should be determined only by the input signal (except for the noise inherent in the amplifier). Under certain circumstances, an amplifier may fall into self-oscillation, generating a periodic waveform that is, in principle, independent of the input signal, although it may become modulated by an input signal or merely added to it. A special case is amplifier drift into a "bottoming" or "cutoff" state, in which the amplifier quickly reaches a constant d-c level at the output. This level is at one end of the output voltage range end-values, where the amplifier becomes blocked. This may be looked on as the extreme case of self-oscillation, "d-c oscillation."

Oscillators

An amplifier that is unstable in the above sense has become an oscillator. To elucidate this, the criteria for oscillation will be discussed. An oscillator is basically an amplifier with *feedback* as shown in Figure 74-19. General aspects of feedback will be discussed later. Here, we will state the criteria for self-oscillation in a feedback system. Let $A(f)$ be the transfer function of the amplifier and $B(f)$ the transfer function of a (generally passive) filter in the feedback path. If we open up the loop at the input to the amplifier, as shown in Figure 74-20, we find a voltage e appearing at the output

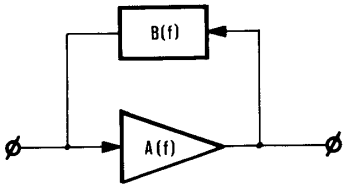


Fig. 74-19. – An amplifier with feedback.

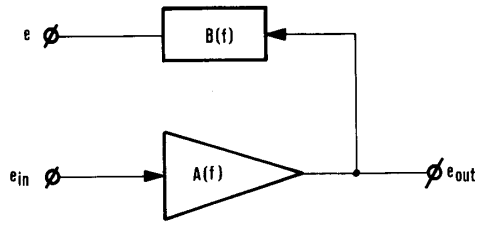


Fig. 74-20. – The feedback loop of Figure 74-19 opened up.

of the feedback filter in response to a sinusoidal input voltage, e_{in}, on the amplifier:

$$e = A(f)\, B(f)\, e_{\text{in}} \qquad (74\text{-}48)^*$$

Here, we presuppose that the effect of the loading of the amplifier input impedance on the feedback filter is included in $B(f)$. If e and e_{in} are equal, that is,

$$A(f)\, B(f) = 1 \qquad (74\text{-}49)$$

we can remove the input source and connect the output from the feedback filter to the input without changing the state of the system. Then the voltage e_{in} will persist at the input of the amplifier after the loop is closed, sustained by the feedback. This means that the circuit oscillates at a frequency (or frequencies) satisfying equation (74-49). This equation, in fact, is the basic oscillation criterion. It actually is a pair of equations, since A and B are complex:

$$\begin{cases} Re(AB) = 1 \\ Im(AB) = 0 \end{cases} \quad \text{or} \quad \begin{cases} |AB| = 1 \\ \angle AB = 2n\pi \end{cases} \qquad (74\text{-}50)^\dagger$$

where n is an arbitrary integer. The latter part of equation (74-50) is more useful for theoretic discussion, although both are equivalent. The case $|AB| = 1$ actually is a special case in which the assumed sinusoidal input voltage would remain unchanged after closing the loop. Equation (74-50) is the criterion for purely sinusoidal oscillation (Barkhausen's oscillation criterion). If $|AB| > 1$ when e_{in} and e are in phase, it is easy to imagine that voltages will be built up in the amplifier after closing the loop, because $e > e_{\text{in}}$. This will continue until the voltages are limited by some non-linearity in the system. If $|AB| \gg 1$, the limits are likely to be set by bottoming and cutoff of the amplifier and the oscillation then will produce a more or less square wave. The non-linear effects then will cause the frequency to deviate from the solution to $\angle AB = 2n\pi$, which will merely be the initial value of the frequency as the oscillation is built up. If $|AB|$ is only slightly larger than 1, the limits will be set by small non-linearities in the input-output characteristic, before bottoming or cutoff is reached. This will make the curve turn on the levels where $|AB|$ is reduced to exactly 1 and the oscillation is approximately sinusoidal. The frequency then will be given approximately by equation (74-50). We now realize that the general criterion for oscillation is

$$|AB| \geq 1 \text{ when } \angle AB = 2n\pi \qquad (74\text{-}51)^\ddagger$$

Obviously, any small disturbance occurring in a closed loop with $|AB| > 1$, when $\angle AB = 2n\pi$, not originally oscillating,

*$A(f)$ and $B(f)$ are two amplifiers in cascade, in this case.

†Re = real part, Im = imaginary part. These are mathematical terms, labeling the components of a complex quantity. The imaginary part is the one having $j = \sqrt{-1}$ as a factor, given its name in earlier days when the concept of the j was a little difficult to visualize. In physical applications, complex notations generally result from (1,1) mappings from the "real world" and therefore the imaginary part is, physically, no less "real" than the real part.

‡I.e., over-all loop amplification greater than or equal to unity when its corresponding input and output signals are in phase.

will cause the build-up of an oscillation. Such a disturbance may be the noise inherent in the circuitry. Hence, we put as stability criterion:

$$\text{Stability} \Rightarrow |AB| < 1 \text{ when } \angle AB = 2n\pi \quad (74\text{-}52)$$

Stability

If a feedback system is to be used as an amplifier, it must fulfill the criterion (equation [74-52]), or else it will function as an oscillator. Actually, this holds for any amplifier even if no intentional feedback circuits are built in; certain feedback paths always exist through parasitic capacitances and through mutual inductance between coupling wires. In the capacitive case, one must aim at reducing such feedback paths as much as possible by keeping input and output circuitry well apart or separated by a grounded shield. In the case of inductive feedback, one must see that wires carrying input and output current do not run too closely or in parallel with each other. Especially, they must not interweave as sketched in Figure 74-21, since this generally results in a serious transformer action. In the case of intentional feedback, one must control the *loop gain* characteristic, i.e., the characteristic of AB, such that $\angle AB$ does not reach 0 or 360° as long as $|AB| \geq 1$. As the values of components can never be exactly specified, one must have a certain *phase margin* $\alpha > 0$: the value of $\angle AB$ must not come closer than α to 0 or ±360° as long as $|AB| \geq 1$.

An amplifier having a high-frequency roll-off of −6 db/octave up to frequencies where $|A| < 1$ usually does not cause much trouble with feedback. Many integrated operational amplifiers, however, begin to roll off with −12 db/octave or more before $|A| < 1$. This usually calls for special measures to keep the circuit stable if feedback is used. Several possibilities exist; careful design of the feedback path is the most obvious; adjustments of the phase characteristics of the amplifier itself, e.g., by using so-called "lead" and "lag" compensations also are useful means. The measure required must be chosen in each case, depending on amplitude and phase characteristic of the amplifier. Op amp manufacturers often give suggestions in their "Application Notes."

Ground Loops

A special problem is posed by the formation of ground loops. Return currents in a ground shield or chassis plate may induce currents in other structures by transformer action and, furthermore, ground connections may interweave, like the input and output wires in Figure 74-21. Such interweaving must be avoided through careful layout of ground and other connections.

A grounded shield, chassis or plate, never has zero resistance. Signal currents in "grounds" may build up sufficiently high voltages over the resistances of the metal to provide a feedback path, as sketched in Figure 74-22. Such a feedback often is sufficient to cause self-oscillation at a high frequency.

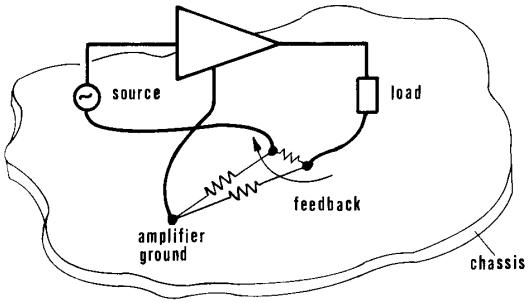


Fig. 74-22.—The resistivity of the chassis metal can give undesired feedback paths.

To avoid trouble with grounds, a general rule is to separate signal ground and power ground (used for the supply of power to active components). There should be only one grounding point for each of these two types, joined together using a heavy wire. That is, all signal ground should be made at one and the same point on the chassis (at least within one stage) and the same goes for the power ground (note that signal current can be superimposed on power current!). Obviously, the use of several ground points can cause interweaving current *paths* in the chassis, even if the wires do not interweave.

Bottoming and Cutoff

A d-c feedback may cause bottoming or cutoff if the "oscillation" criterion (equation [74-51]) is fulfilled at zero frequency. A d-c voltage is thus established at the output, at a level at which the amplifier ceases to work as such, as active components become saturated or blocked. This must be considered when designing d-c feedback paths.

A special problem here is the possibility of *thermal feedback*. Power dissipation in a transistor causes heating, which lowers the base-emitter voltage (negative temperature coefficient) and increases leakage currents. This may cause an *increase* in power dissipation, causing further changes in transistor parameters, until the transistor saturates or is burned out. This is called *"thermal runaway"* and is a feedback phenomenon in which electrical and thermal effects are coupled together in a loop. Very-low-frequency oscillations are possible in such a case, e.g., if nearby components are thermally coupled, but generally it results in a d-c drift to bottoming or cutoff. Thermal runaway is prevented by sufficient cooling of the device ("thermal grounding"), by choosing appropriate working points of the transistors and by providing counteracting (stabilizing) d-c

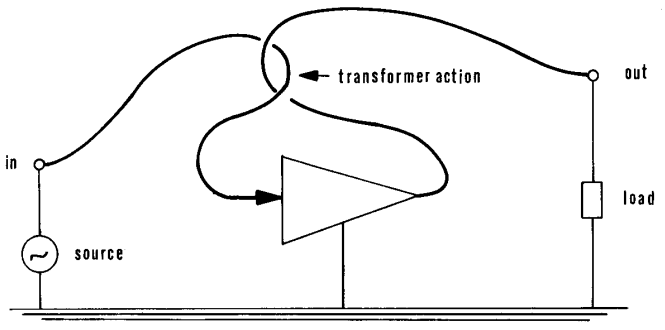


Fig. 74-21.—Interweaving wires give transformer action that can result in an undesired feedback path.

feedback. A thermal-to-electric stabilizing feedback may be realized using a thermistor in thermal contact with the transistor case.

Reliability and Systematic Errors

In the preceding part we discussed stability in the sense that an amplifier can be kept from falling into self-oscillation. Here, "stability" in a different sense will be discussed – "long-term stability" or *reliability*, related to possibilities of deterioration of performance through aging. Before getting to this, we will discuss how the performance of a circuit can be affected by deviations from nominal values of component parameters, and how the sensitivity to parameter variations can be reduced through feedback.

Passive Components – Standard Values and Tolerances

Two factors generally causing deviations of component values in the realized circuit, from those theoretically calculated, are that components are commercially available only with certain nominal standard values and that, furthermore, the actual values differ from the nominal ones within specified tolerance ranges. Common tolerances for resistors and capacitors will be given below; of course, closer tolerances are available at higher prices, and economy calls for design of a circuit that is sufficiently insensitive to variations in component values, so that components with wider tolerance ranges can be used. A circuit should tolerate the variations in component values from a theoretic to a nearest standard value, so that single standard-valued components can be used.

Standard values of resistors in the so-called "R12-series," having 12 values per decade, are, in principle, determined by

$$R = 10^{\frac{n}{12}+k} \qquad (74\text{-}53)$$

where n and k are integers; $0 \leq n \leq 11$, k is any integer and $10^{n/12}$ is approximated to one decimal place only. In practice, a few values are shifted from the theoretic value by one unit in the decimal place for better interlace of tolerance ranges in different series, such as the "R24-series" and "R6-series" having 24 and 6 values per decade, respectively. The most common series is the above-defined "R12-series" with the following $10^{n/12}$ values:

1.0	2.2	4.7*
1.2	2.7*	5.6
1.5	3.3*	6.8
1.8	3.9*	8.2*

where the values marked by * are slightly shifted, as mentioned above. This series generally has a tolerance of 10%, whereas the "R24-series" generally has 5% tolerance and the "R6-series" has 20% tolerance. Other than standard values are made, however, especially for precision resistors with a tolerance of 1% and better.

Modern capacitors, too, usually are made according to the same standard rules as those for resistors, generally as a "12-series" (equation [74-53]). But here the flora of nonstandard values is rich, especially for electrolytic capacitors. Common tolerances are 20%, 10% and 5% for normal capacitors and up to 50% or 100% or even more for some types of electrolytic capacitors.

Temperature Effects

Resistors and capacitors change their values with temperature, generally in a fairly linear fashion. Electrolytic capacitors, however, can have a pronounced non-linear temperature characteristic. Temperature coefficients or characteristics usually are given by the manufacturers. Resistors and normal capacitors may generally change within 1% and 10% for a temperature change of 100° C, with positive or negative temperature coefficients, depending on the value, material and construction. Electrolytic capacitors may change much more, as may cheap non-electrolytic capacitors.

Precision components are available with a much lower temperature coefficient, even with a differential coefficient of zero value at a suitable working temperature.

Aging, Failure Rate

Component values generally change with time, although, on the average, at a very slow rate. Individual components may, however, change quite rapidly; whether they do or not is a question of probability. We define a failure as the event when the value of a component has deviated from its nominal value more than is tolerated, in view of the performance of the circuit employing it. This meaning of failure then depends on the application. For practical reasons, one generally gives data for a change exceeding a certain fixed percentage of the nominal value. If the probabilty that a component, put into operation at time zero, will not fail before time t is $R(t)$, the probability that it will fail in the time interval from t to $(t + \Delta t)$ is, for small Δt:*

$$p(t)\Delta t = \Delta t \frac{d}{dt}[1 - R(t)] = -R'(t)\Delta t \qquad (74\text{-}54)^*$$

where $p(t)$ is the failure density function. The rule for combination of conditional probabilities states that this can also be written as the product of the probability $R(t)$, that component does not fail up to time t, with the probability $[q(t)\Delta t]$, that it fails in the interval from t to $(t + \Delta t)$ under the (conditional) assumption that it did not fail up to time t:

$$p(t)\Delta t = R(t)\, q(t)\, \Delta t \qquad (74\text{-}55)^\dagger$$

The quantity $q(t)$ is not a failure density function in the usual sense but a *conditional* failure density function called the *failure rate*. A typical course of $q(t)$ is sketched in Figure 74-23. This function may be thought of as the quotient between the number of components failing within a time interval of unit length and the number of components not failing until that interval begins, having started with a large number of components.

As sketched in Figure 74-23, the failure rate generally has a broad region at a constant level, which we denote with q_o. Then, in this region:

$$q_o = -\frac{R'(t)}{R(t)} \qquad (74\text{-}56)$$

from equations (74-54) and (74-55). Integrating equation (74-56) gives

$$R(t) = R\, e^{-q_o(t-t_1)} \qquad (74\text{-}57)$$

where R_o is the probability that the component lasts until the time t_1, where the constant level begins. q_o is a small number, much less than unity, and if we have a combination of n components in a circuit, each component having the same $q(t)$, the failure rate $q_n(t)$ for the combined circuit can be found as follows. The conditional probability that a single component does not fail in an interval of length Δt, between t_1 and t_2 (see Fig. 74-23), is $(1 - q_o\Delta t)$. The conditional probability that none of the n components fails is the

*Δt, "delta t," symbolizes a small deviation from t; the probability distribution function for failure is $[1 - R(t)]$ and the corresponding probability density function is $p(t)$. Equation (74-54) actually is the definition of their interrelation.

†Probability, that it lasts until t, times probability that it then fails within time Δt.

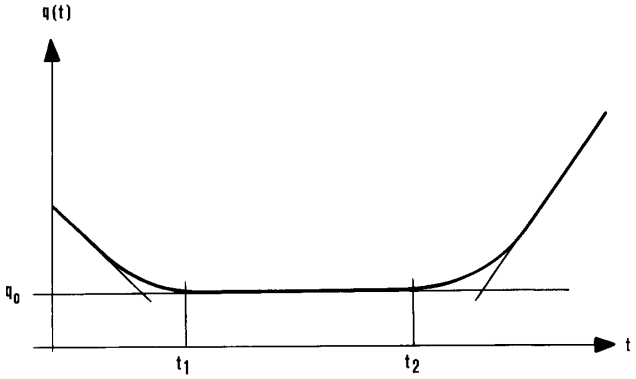


Fig. 74-23. – Sketch of failure rate as a function of time.

n:th power of this. The conditional probability that *at least* one component fails then is

$$q_{on}(t) = 1 - (1 - q_o\Delta t)^n \approx nq_o\Delta t \qquad (74\text{-}58)^*$$

as q_o is small, which is also assumed for Δt. Hence, the probability R_n that the circuit *lasts* until time t, between t_1 and t_2, is

$$R_n(t) = R_{on}e^{-nq_o(t-t_1)} \qquad (74\text{-}59)$$

from equation (74-57). We see that $R_n(t)$ not only falls *exponentially* with time but also with the number of components.

Generally, $q(t)$ is different for the different components in the circuit, but, in general, it still holds that the risk that the circuit fails before a certain time grows not only with time but also with the number of components. In any case, the components with the highest failure rate have the largest influence on this risk – "the weakest link in a chain."

Because $q(t)$ falls off at the beginning, before time t_1 in Figure 74-22, components for precision applications sometimes are pre-aged so that they will have passed their time t_1 and reach the lowest q level without failure before being put into the circuit. $q(t)$ is also largely influenced by such operating conditions as temperature, humidity, mechanical stress, etc.

Semiconductor Components

Transistors and diodes have aging properties similar to those of passive components. Often the *1/f* noise of a transistor increases sharply before the transistor fails with respect to its operating parameters.

The spread of some parameter values of semiconductors usually is several magnitudes larger than for passive components. The base-to-collector current gain β of a standard transistor may have a relation between maximal and minimal values of 2 to 5. The leakage currents may differ even more – for both transistors and diodes. The base-to-emitter voltage, although generally showing less spread, may have a maximal value twice the minimal; this is also the case for the forward voltage drop of a diode at a given current.

Variation of parameters with temperature can also be substantial. Leakage currents follow an exponential temperature law, doubling the value for approximately every 12° C increase in temperature. The base-to-emitter voltage is reduced by roughly 2 mV/°C temperature rise and β may increase by the order of 0.5% per °C rise.

It is therefore of special importance to design circuits that can accept large variations in active-component parameters. They should never be designed according to measured parameters of individual transistors, nor should they have trimmer potentiometers to adjust for deviations in parameter values when transistors must be exchanged later. In rare cases and for special purposes, such designs nonetheless can be accepted; it often may be necessary, for example, to put in a trimmer for adjustment of the offset of a d-c amplifier.

*If $|\epsilon| \ll 1$, one has $(1 + \epsilon)^n \approx 1 + n\epsilon$.

Reduction of Sensitivity to Changes in Component Parameters

A certain insensitivity can be achieved by choosing a suitable d-c operating point for each transistor. A more general method, however, is the use of *feedback*, as discussed in the following general terms.

Feedback was touched on in the discussion on stability, but not as to its effect on input and output signals. Here, we will see the effect of a feedback loop so arranged that the output of the feedback filter adds to the input signal, as is sketched in Figure 74-24. The stability criterion for this arrangement is independent of e_i; thus, when considering stability, we may put $e_i = 0$ and make sure that the criteria previously given are fulfilled.

The equations describing the circuit in Figure 74-24 are, for sinusoidal signals,

$$\begin{cases} e_f = Be_o & (74\text{-}60) \\ e_o = A(e_i + e_f) & (74\text{-}61) \end{cases}$$

Solving for e_o, we find

$$e_o = e_i\frac{A}{1 - AB} \qquad (74\text{-}62)$$

where stability requires $|AB| < 1$ when $\angle AB = 2n\pi$. Since generally $|AB| > 1$ within the passband of A, one requires $\angle AB \neq 2n\pi$ in that region, or, equivalently, $\cos \angle AB < 1$. The extreme case of this condition is $\cos \angle AB = -1$, meaning that AB is real and *negative*. This offers the greatest safety against self-oscillation and therefore generally is sought within the passband. This is called *negative feedback* and is employed to *counteract* changes in the amplifier (positive feedback, where AB is real and positive, generally gives an oscillator, except when $AB < 1$, enhancing changes in the amplifier). The terms "negative" and "positive" feedback may be generalized to the cases $\cos \angle AB < 0$ and $\cos \angle AB > 0$, respectively.

The most critical components are the transistors in "A" ("B" generally is a passive network), with their wide spread in characteristics and their temperature dependence. Therefore, the value of A may vary considerably with temperature and time or if a new transistor is used to replace an original, faulty one. If, however, $|A|$ is made very large in its

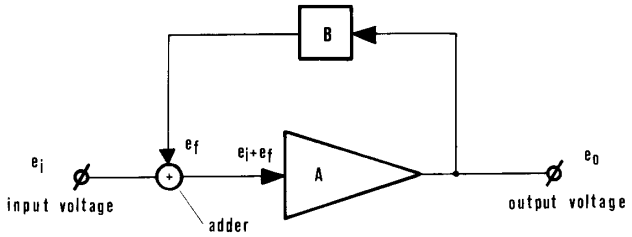


Fig. 74-24. – General case of an amplifier with feedback.

passband, so that also $|AB| \gg 1$, we get from equation (74-62):

$$e_o \approx -\frac{e_i}{B} \tag{74-63}$$

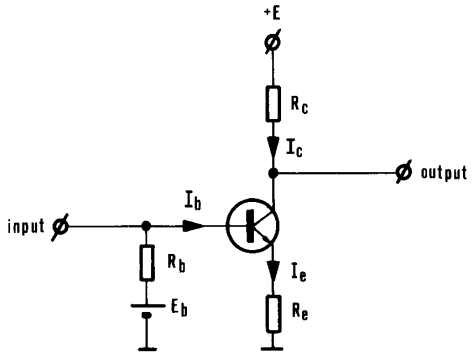
or an over-all amplification of $-1/B$, independent of A. Thus, the circuit is, through feedback, made largely insensitive to alterations in transistor parameters. Equation (74-63) clearly requires $B < 1$ in the passband, from which we conclude that a passive circuit generally is sufficient as a feedback filter.

The condition for reduction of sensitivity to variations in A, according to equation (74-63), is that $|AB| \gg 1$. If only signal feedback is employed, A occasionally may change so much that this condition is not fulfilled, due, for example, to shifts in the d-c operating points of the transistors. To counteract this, the feedback must also be active at d-c. For practical reasons, this usually calls for separate signal and d-c feedback paths. In most cases, each individual transistor stage has its own d-c feedback, which may simply be realized by a resistor, in the emitter connection shown below in an example. Further d-c feedback paths over several stages may be employed as well, together with one or more signal feedback paths, over the whole amplifier or parts of it. As is discussed in Chapter 72, the signal feedback may also be used to tailor the over-all characteristics of the amplifier in a way largely independent of A. Also, input and output characteristics may be altered through feedback, as well as non-linearities in the amplifier, which will be discussed later.

As an example of very simple d-c feedback in a single transistor stage, we will use an emitter-resistor circuit such as is shown in Figure 74-25. Since circuit analysis is not the subject of this chapter, an oversimplified model of the transistor will be used, neglecting leakage currents, base-emitter voltage and input and output resistances of the transistor. We will only show how the influence of β, the base-to-collector amplification, is reduced through the use of R_e. For d-c levels, the input is the base d-c drive E_b and the output is the collector d-c current I_c; they are related as

$$I_c = \beta I_b \tag{74-64}$$

Fig. 74-25. – Use of an emitter resistor (*Re*) for stabilization of the collector d-c current (I_c) against variations in the current amplification of the transistor.



Without R_e (emitter directly grounded), we would have

$$I_b = \frac{E_b}{R_b} \tag{74-65}$$

and

$$I_c = \frac{\beta E_b}{R_b} \tag{74-66}$$

varying proportionally with β. I_b is determined by the voltage developed over R_b, which, in equations (74-64)–(74-66), is E_b. When R_e is introduced, this voltage becomes $E_b - R_e I_e$, where I_e is the d-c emitter current, $I_e = (1 + \beta)I_b$. Generally, $\beta \gg 1$ and we can put $I_e = I_c$. The reduction of the voltage at R_b by an amount proportional to the output I_c means negative feedback, with R_b as the "adder" in Figure 74-24. We then have

$$B = -R_e$$

and, from equation (74-62):

$$I_c = E_b \frac{\beta/R_b}{1 + \beta R_e/R_b} \tag{74-67}$$

as is also verified by direct calculation in Figure 74-25. If $\beta R_e \gg R_b$, we get

$$I_c = \frac{E_b}{R_e} \tag{74-68}$$

independently of β. The employment of an emitter resistor is an important "trick" to tolerate the wide spread in β and other transistor parameters. The emitter resistor, however, means a reduction of signal amplification, which generally is overcome by bypassing the signal current in a large capacitor in parallel with R_e. This introduces a low-frequency roll-off and hence cannot be used in amplifiers with d-c response. In such amplifiers, either the reduction of amplification in the single stage is accepted or other means of feedback for working-point stabilization can be used (any d-c feedback in a d-c amplifier acts on the signal as well; it then is generally advisable to use a feedback over as many stages as possible rather than in each single stage, so that "d-c" and "signal" feedback cooperate).

Redundant Amplifiers

With high or excessive "forward" amplification and subsequent reduction of the over-all gain to a desired level through feedback, the sensitivity of the amplifier to certain component parameters is seen to be reduced; hence, the

failure rate of the amplifier is reduced. Another way to reduce failure rate is to use a redundant combination of amplifiers, employing several in parallel, so that the still-functioning ones can "do the job" if one fails. This may be an expensive way to reduce failure rate, although it is quite effective and sometimes used where high precision and long lifetime are required or failure would be catastrophic. Still another way to provide this insurance is to employ a second, "idling" amplifier. Normally, it is not operating but is switched in as a replacement whenever the normally operating amplifier fails. This "replacement" is automatically controlled by a means for the detection of an excessive change in some important characteristic.

Stability

In the discussion on stability it was mentioned that a certain "phase margin" is required in practice. This now should stand out clearer in view of tolerances. If no "phase margin" is allowed for, a change of a component parameter may shift the phase characteristic so that the stability criterion no longer is fulfilled and the amplifier starts to oscillate.

Systematic Error

A change in some amplifier characteristic due to aging, temperature effects or the exchange of a damaged component, as discussed above, clearly can lead to a systematic error (an error that is stable, measurable and generally can be compensated for) in a measurement.

Another important source of systematic errors in a d-c amplifier is the offset voltage. Generally, the output voltage e_o for a d-c amplifier can be written as a function of input voltage e_i, as

$$e_o = A(e_i + E_{off}) \qquad (74\text{-}69)$$

where E_{off} is a d-c voltage arising from d-c shifts and levels within the amplifier. E_{off} is largely dependent on component parameters and usually changes with temperature, aging, etc. To keep a proportionality between e_o and e_i, it is necessary to make E_{off} zero. In some cases, this may be done through feedback, e.g., common-mode feedback in a differential amplifier (see below). Often an adjustable shift of a d-c level within the amplifier is introduced, so that E_{off} can be adjusted to zero (temperature compensation also may be employed so that the adjustment need not be repeated after a change in temperature). Another possibility is to use a chopper amplifier (see discussion on noise), in which a corresponding a-c signal is amplified independently of E_{off}.

Another systematic error can be that introduced by variations in the power supply voltage. Such variations may alter operating points and, therefore, amplifier characteristics, as well as the d-c offset. D-c feedback and closely regulated supply voltages, which must be independent of current drain, are means to counteract this. Another problem may be power line hum, introduced into the circuit via the power supply. This hum may add to the signal in the amplifier or, which is worse, modulate the signal through variations of transistor operating points at the mains frequency. The appropriate therapy generally is better filtering in the power supply and in the d-c power paths within the amplifier. In critical cases, one may increase the power line frequency (e.g., to 400 Hz) to make filtering easier, or one may use separate supply-voltage stabilizers for each amplifier, or each group of amplifiers, fed from a common prestabilized supply.

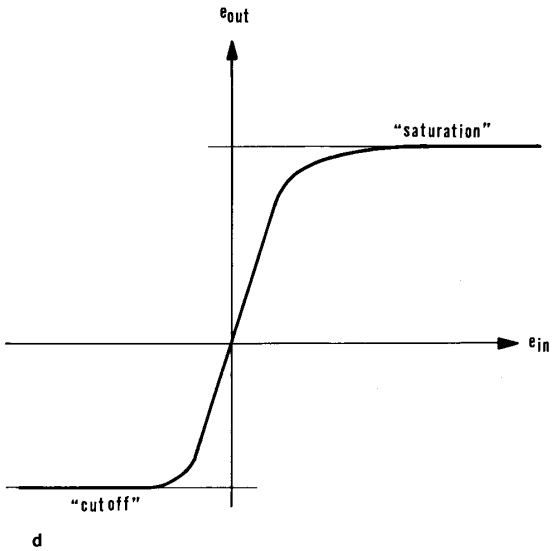
For a differential amplifier, which in the ideal case gives off a voltage proportional to the difference $(e_1 - e_2)$ between its two input voltages, e_1 and e_2, another systematic error arises. The real differential amplifier also still has a certain contribution at the output from the *sum* $(e_1 + e_2)$ of the input voltages. This contribution should be as low as possible. A measure of quality of the amplifier in this respect is the *common-mode rejection ratio*, CMRR, defined as the quotient between the amplifications (voltage gains) of the difference signal $(e_1 - e_2)$ and the mean-value signal $(e_1 + e_2)/2$ (called the common-mode signal), respectively. A differential amplifier should be symmetric in its first stages (see Chapter 71). Deviations from symmetry allow part of the common-mode signal to pass. To make the CMRR high, one should use matched (selected, nearly identical) transistors to increase operational symmetry. A feedback loop, acting only on the common-mode signal, brings further improvement. A high common-mode signal and a small difference-signal applied together at the inputs of a differential amplifier with insufficient CMRR can result in an output signal that deviates considerably from the input difference signal. In such cases, the difference signal may be greatly misjudged.

Deviations from nominal values in amplifier characteristics, caused by aging, abnormal temperature, etc., can result in similar or other types of mesurement errors.

Non-Linear Effects

The over-all performance characteristic of a d-c amplifier at low frequencies generally is linear only within a certain range of output and corresponding input voltages. A typical characteristic is sketched in Figure 74-26. The characteristic generally approaches constant levels for high $|e_{in}|$ (in some cases it may, for example, turn and become reversed). One extreme level is called saturation, the other cutoff. In saturation, a transistor acts as a short circuit; when cutoff, it is non-conducting. At the limiting levels of an amplifier, the states usually are different for the different transistors in the total amplifier. The terms "saturation" and "cutoff" may then, as used for describing the characteristic of the amplifier taken as a whole, in effect refer to the state of the output transistor or output "stage."

Fig. 74-26.—Typical amplifier input-output characteristic (at low frequencies).



The characteristic usually is smooth and therefore can be described by a McLaurin series (power series) in terms of e_{in}:

$$e_{out} = A_1 e_{in} + A_2 e_{in}^2 + A_3 e_{in}^3 + A_4 e_{in}^4 + \ldots \quad (74\text{-}70)^*$$

at low frequencies, where the coefficients are constant and the d-c offset, which adds a constant term, is neglected. High-frequency behavior in the presence of non-linearities will be discussed briefly later.

If the input signal is sinusoidal, $e_{in} = E \sin 2\pi ft$, the series (equation [74-70]) gives

$$e_{out} = A_1 E \sin 2\pi ft + \frac{1}{2} A_2 E^2 (1 - \cos 4\pi ft) + \frac{1}{4} A_3 E^3 (3 \sin 2\pi ft - \sin 6\pi ft) + \frac{1}{8} A_4 E^4 (3 - 4 \cos 4\pi ft + \cos 8\pi ft) + \ldots \quad (74\text{-}71)^\dagger$$

i.e., a single sinusoid gives rise to *harmonics* at the output; their sum actually is the Fourier series for the distorted sinusoid appearing at the output. This may cause considerable disturbance, as the harmonics of the lower spectral components in the input signal add to higher spectral components. Therefore, it may not be possible to distinguish certain original components from the harmonics generated in the amplifier when looking at the output. Even worse may be that signal components at different frequencies mix to form components at sums and differences between integer multiples of the individual frequencies. If the spectrum of the input signal has one single peak, harmonics of this peak result in the appearance of a series of peaks at integer multiples of the basic frequency in the spectrum of the output signal—when there is too much non-linearity within the actual range of the output signal. In spectral studies, one therefore must keep in mind that such "secondary" peaks may be artifacts. The suspicion arises whenever peaks appear at integer multiples of the basic frequency.

The coefficients in equation (74-70) usually are falling in magnitude, $A_1 > A_2 > \ldots$ Linearity, therefore, is greatly improved if we can get rid of the second-power term, especially as this term is an even function of e_{in}, whereas the linear term and the third-power term both are odd functions of e_{in}. The A_2 term actually can be removed by using a suitable operating point. Due to the general appearance of the input-output characteristic, it must have at least one point of inflection between the asymptotic levels. If the characteristic is so shifted (by suitable d-c shifts within the amplifier) that a point of inflection falls at the origin of the e_{out}, e_{in} coordinate system, the second-power term by definition is zero:

$$e_{out} = A_1 e_{in} + A_3 e_{in}^3 + A_4 e_{in}^4 + \ldots \quad (74\text{-}72)$$

Thereby, also the second harmonic in equation (74-71) is reduced to the generally much lower contribution from the A_4 term and following even-order terms. Further improvements can be achieved through feedback and by tailoring the output range so that its asymptotic levels lie far off from the desired range of output signals. The former method will be discussed below.

Reduction of Non-Linearity through Feedback

An amplifier with a non-linearity can be represented by two ideal amplifiers with a source of harmonics connected in between, at the site of the non-linearity; this is sketched in Figure 74-27. In this figure, a feedback path also is introduced. As was seen in the discussion on reliability, the contribution of a sinusoidal voltage e_{in} to e_{out} is:

$$e_{out\,1} = e_{in} \frac{A_1 A_2}{1 - A_1 A_2 B} \quad (74\text{-}73)^*$$

whereas the contribution of each sinusoidal component e_{hi}, $i = 1,2 \ldots$, of e_h in Figure 74-27 is

$$e_{out\,2i} = e_{hi} \frac{A_2}{1 - A_1 A_2 B} \quad (74\text{-}74)^\dagger$$

as here the gain in the feedback path is $A_1 B$. The relation between the two contributions is

$$\frac{e_{out\,2i}}{e_{out\,1}} = \frac{e_{hi}}{A_1 e_{in}} \quad (74\text{-}75)$$

which should be as low as possible. Now, if $A_1 A_2$ is made high, the desired over-all amplification is determined by B:

$$e_{out\,1} = -\frac{e_{in}}{B} \quad (74\text{-}76)$$

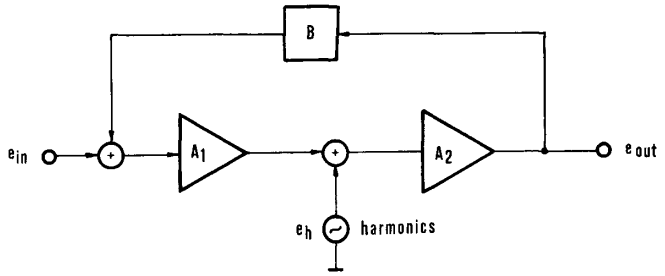
(see above) and we therefore can reduce this influence of the non-linearity by making A_1 as large as possible (see equation [74-75]) and return the over-all amplification to the desired value with B.

Measure of Non-Linearity

As a measure of the non-linearity of an amplifier, one uses the *harmonic distortion factor*, defined as:

$$k = \frac{\sqrt{E_2^2 + E_3^2 + E_4^2 + \ldots}}{\sqrt{E_1^2 + E_2^2 + E_3^2 + \ldots}} = \sqrt{1 - \frac{E_1^2}{2E_o^2}} \quad (74\text{-}77)$$

Fig. 74-27. – Use of feedback for linearization.



*A large class of mathematical functions can be expressed as a (generally) infinite sum of powers of the independent variable. More generally, a constant term should be included, here assumed to be zero.

†Develop the powers of sin 2 πft.

*An amplifier $A_1 A_2$ with a feedback B.

†An amplifier A_2 with a feedback $A_1 B$.

if the output signal, corresponding to a sinusoidal input $E \sin 2\pi ft$, is (see equation [74-71])

$$e_{out} = E_1 \sin 2\pi ft + E_2 \sin 4\pi ft + E_3 \sin 6\pi ft + \ldots \quad (74\text{-}78)^*$$

and the rms value of e_{out} is E_o. k can be calculated from measurements of the total rms value E_o and the rms value $E_1/\sqrt{2}$ of the filtered-out basic component obtained via a suitable filter in series with the output, removing harmonics. It generally varies with both amplitude E and frequency f of the input signal.

Non-Linearity at Higher Frequencies

The above discussion was based on the properties of a d-c amplifier at low frequencies. It also holds for a broadband amplifier with small phase shift within the part of the passband containing the main part of the harmonics.

If phase shift occurs, the input-output characteristic shows a kind of hysteresis loop. Hysteresis, in a general sense, is present when the input-output relation follows different paths, in the diagram of output quantity vs. input quantity for increasing and decreasing input signal (see Fig. 74-28). For a linear amplifier with a sinusoidal input, this appears as an ellipse in the input-output diagram. This "linear hysteresis" is not a harmonic distortion! When non-linear effects exist, the above appears as a deformed ellipse or a closed loop of any shape.

A qualitative study of harmonic distortion with phase shift, by checking the form of the loop in the input-output diagram, is admissible only when the input signal is sinusoidal. Other waveforms do not generate ellipses even with linear amplifiers, since the harmonics then already present in the input signal undergo different phase shifts. Figure 74-28 shows three cases: a linear circuit with a sinusoidal input, a triangular input and a non-linear circuit with a sinusoidal input.

A case considerably worse occurs when the input-output relation does not immediately form a loop. It may not, after one cycle, have returned to the starting point and the subsequent curves generate a spiral that only asymptotically approaches a closed loop. This is a case of severe non-linear distortion with memory effect, so that the state of the circuit also depends on the past history. Such a case is also shown in the fourth diagram of Figure 74-28.

Blocking

Another effect of non-linear origin, which can cause serious disturbances, is that a transient, appearing at the input, may drive the input transistor of the amplifier into cutoff. The input capacitance to the amplifier becomes charged by the transient and the input resistance raised due to the cutoff condition. The latter effect may increase the time constant of input capacitance and resistance by orders of magnitude, which, in turn, can cause the transistor to remain in the cutoff state for a relatively long period. During this period, the amplifier is blocked and does not pass signals. The cure is to reduce input capacitance, reduce the input voltage or limit the input signal, e.g., with diodes, so that transient spikes are clipped.

Another kind of blocking occurs when the input transistor is driven to saturation. Return to normal operating level

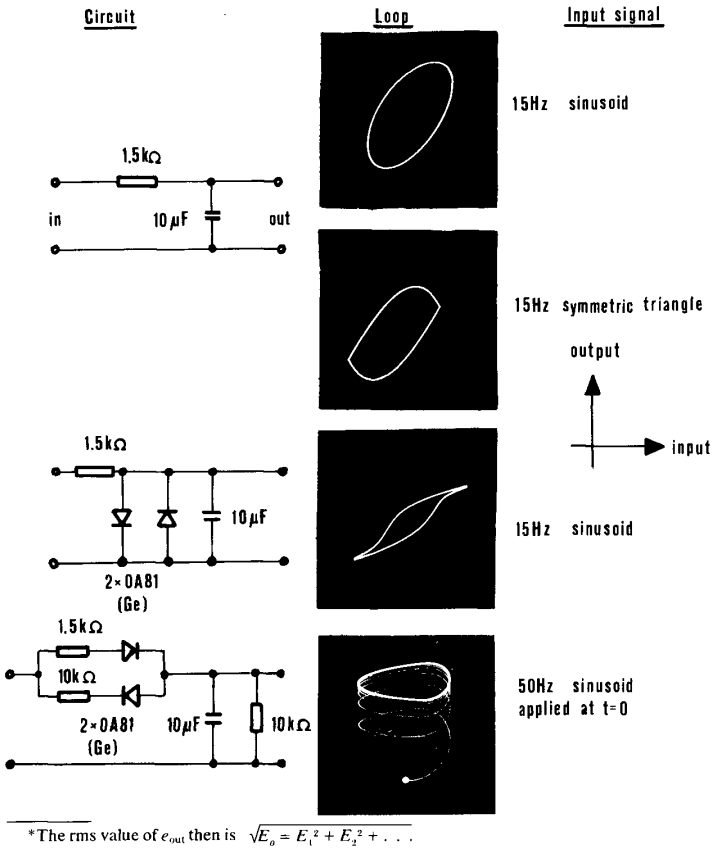


Fig. 74-28. – Input-output relations at higher frequencies for a linear circuit (*top*), and non-linear circuits. (Photographed from oscilloscope tracings.)

*The rms value of e_{out} then is $\sqrt{E_o = E_1^2 + E_2^2 + \ldots}$

from saturation requires a certain time, as the accumulated base charge in the transistor must be removed. However, the time period of this blocking is, generally, negligible in most biomedical applications.

SUMMARY

It has been shown that non-linear effects in the amplifier, apart from causing deformations of the waveform of the signal, may distort and add artifacts to the signal spectrum. These effects can be reduced by choosing appropriate d-c working points for the transistors and by applying feedback. A special effect of the non-linearity caused by cutoff of a transistor stage is that it may become blocked for certain time periods following large transient spikes.

Input and Output Impedances

This part will cover the effects of signal source impedance and the loading impedance connected to the amplifier at the output. The input impedance to an amplifier loads the signal source; how this will affect amplifier behavior depends on the source impedance. The loading impedance connected to the amplifier output may influence the output signal, depending on the output impedance of the amplifier. The term "loading" refers to the consumption of power that may be delivered by the source or the amplifier.

THE FILTERING EFFECT OF SOURCE AND LOAD IMPEDANCES

Every signal source has a source impedance. As is discussed in Chapter 69, the source can be represented in either of two equivalent ways, as sketched in Figure 74-29, using an ideal source (voltage source of zero impedance or current source of infinite impedance) and an "output" impedance. Here, we will use the voltage source representation. Z_s is the *source impedance*, which cannot be circumvented, i.e., the current necessarily flows through Z_s and cannot be drawn directly from the ideal source e. If now the source is loaded by an impedance Z_l, connected across the output, as shown in Figure 74-30, we get the output voltage

$$e_o = \frac{Z_l}{Z_s + Z_l} e \qquad (74\text{-}79)$$

if e is sinusoidal. If it is not sinusoidal, we get the same relation between the *spectra* of the signals instead, as discussed above. For simplicity, sinusoidal signals will be discussed here. The current becomes:

$$i_o = \frac{e}{Z_s + Z_l} \qquad (74\text{-}80)$$

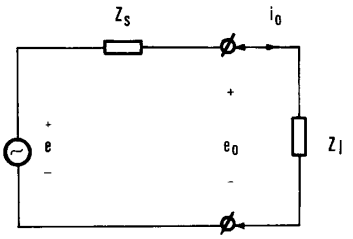


Fig. 74-30. – Effect of loading of the source.

and the power developed in the load

$$P_o = \mathrm{Re}(e_o i_o^*) = \left|\frac{e}{Z_s + Z_l}\right|^2 \mathrm{Re} Z_l \qquad (74\text{-}81)^*$$

where *Re* denotes the real part of the complex quantity following it. Writing $Z_l = R_l + jX_l$ and $Z_s = R_s + jX_s$, we find

$$P_o = \frac{|e|^2 R_l}{(R_s + R_l)^2 + (X_s + X_l)^2} \qquad (74\text{-}82)$$

Impedance matching is defined as the situation in which the power P_o in the load is maximum. Obviously, we should first make $X_l = -X_s$ in equation (74-82), removing the second set of parentheses in the denominator. Differentiating the remaining expression with respect to R_l, one finds a maximum for $R_s = R_l$. This is illustrated in Figure 74-31, where the course for negative R_l is drawn by a dashed curve; as such an R_l cannot be realized with passive components (it is possible, though, with active circuits). Hence, maximal power is obtained when:

$$\textit{matching:} \begin{cases} R_l = R_s \\ X_l = -X_s \end{cases} \text{ or } Z_l = Z_s^* \qquad (74\text{-}83)^\dagger$$

where Z_s^* is the complex conjugate of Z_s. If this condition is fulfilled, the load impedance is said to be *matched* to the source.

If one wants to measure the voltage generated in the source, however, matching is not the ideal situation. From equation (74-79) we see that, in order to obtain $e_o = e$, we should have $|Z_l| \gg |Z_s|$. This calls for the use of high-input impedance amplifiers in many biomedical applications, where the $|Z_s|$ itself often is high.

If a current measurement is desired, extraction of the current delivered by the ideal current source in the alternative representation in Figure 74-29, the load impedance $|Z_l|$ should fulfill $|Z_l| \ll |Z_s|$. This rarely is required in biomedical applications.

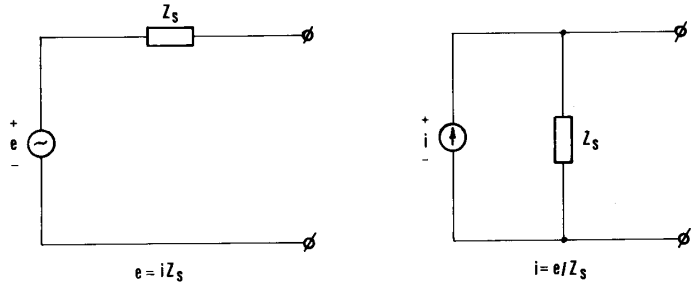


Fig. 74-29. – Alternative, equivalent representation of a signal source: ideal voltage source, with output impedance (*left*) and ideal current source with output impedance (*right*).

*The complex conjugate of a complex number $z = x + jy$ is $z^* = x - jy$. This gives $zz^* = |z|^2$.

†See footnote for equation (74-81).

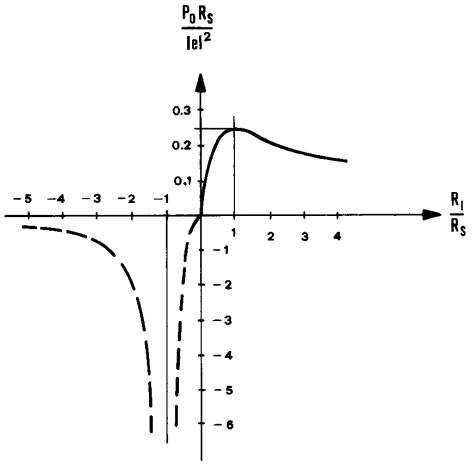


Fig. 74-31. – Variation of load power with load resistance (assuming matched reactances). (The dashed curve holds for negative load resistance as can be realized only with special active circuits.)

Input Characteristic of the Amplifier

The load on the signal source is caused by the total input impedance of the amplifier (including parasitic contributions such as cable capacitance, cable inductance and leakage resistances). This impedance should be adjusted to the source impedance in one of the three ways mentioned above (maximal power, maximal voltage or maximal current) – in biomedical applications generally for maximal voltage, i.e., $|Z_l| \gg |Z_s|$. If this is not possible for all relevant frequencies, i.e., frequencies contained in the main part of the spectrum of the source signal, the latter signal will be *filtered* already at the input to the amplifier, according to equation (74-79). If the impedances Z_s and Z_l are known, compensation can be made for this effect, but Z_s often is not known in sufficient detail for such a compensation. On the one hand, biomedical signals generally are of a low-frequency nature, making the adjustment of the input impedance easier. On the other hand, source impedances often are so high that it is less easy to realize a proper input impedance. Biomedical amplifiers, therefore, often are made with the highest possible input impedances (for voltage measurement) in order to accommodate many types of input transducers, electrodes, detectors, etc.

Output Characteristics

The discussion above, in general, also applies to the output of the amplifier, which can be modeled by a source as in Figure 74-29, subject to the output load attached to the amplifier. Fortunately, the practical problems are much smaller here. It is useful to make the output impedance of the amplifier as low as possible (for voltage measurement) to allow for as wide a variety of loads as possible.

A Special Kind of "Matching" for Constant Amplitude Response

If $|Z_l| \gg |Z_s|$ cannot be enforced (measuring voltage) all over the relevant frequency band, things are relatively good if e_o/e in equation (74-79) is independent of frequency. Putting e_o/e = constant, we find the condition:

$$\frac{X_l}{R_l} = \frac{X_s}{R_s} \tag{74-84}$$

or $\underline{/Z_s} = \underline{/Z_l}$. This often is easier to realize than $|Z_l| \gg |Z_s|$ and gives

$$e_o = \frac{R_l}{R_l + R_s} e \tag{74-85}$$

Note, however, that Z_s must be known in sufficient detail, which limits the applicability of this kind of "matching" in biomedical measurements. Actually, it often is impossible to use it, as Z_s may vary with time and from one application to another.

The Use of Feedback

Input impedances can be raised and output impedances lowered by feedback. In the case of negative-voltage feedback, we subtract from the source voltage another voltage that is proportional to the output voltage. For the input impedance, the situation then can be described as in Figure 74-32, where we find, neglecting load effects on the output,

$$\begin{cases} i_1 = \dfrac{e_o}{AZ_{\text{in}}} & (74\text{-}86) \\ e_1 = e_o\left(k + \dfrac{1}{A}\right) & (74\text{-}87) \end{cases}$$

Here, k is a constant. Hence, the effective input impedance to the amplifier with feedback is:

$$Z_{\text{in, eff}} = \frac{e_1}{i_1} = (1 + kA)Z_{\text{in}} \tag{74-88}$$

which is larger than Z_{in}.

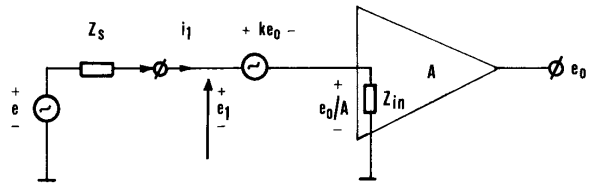
For the output impedance, we may use the representation of Figure 74-33, where, for simplicity, we neglect load effects at the input. Here, we find

$$e_o = e\frac{AZ_l}{Z_o + (1 + kA)Z_l} = e\frac{A}{1 + kA}\frac{Z_l}{Z_l + \dfrac{Z_o}{1 + Ak}} \tag{74-89}$$

which can be represented, as in Figure 74-34, using an effective output impedance

$$Z_{o,\text{ eff}} = \frac{Z_o}{1 + kA} \tag{74-90}$$

Fig. 74-32. – Effect of feedback on input impedance; ke_0 represents the feedback voltage added at the input (see Fig. 74-24).



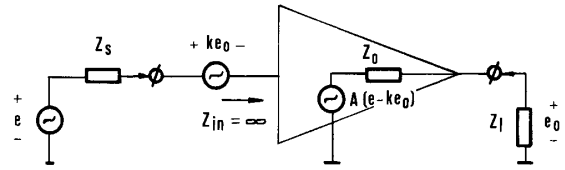


Fig. 74-33. – Effect of feedback on the output impedance.

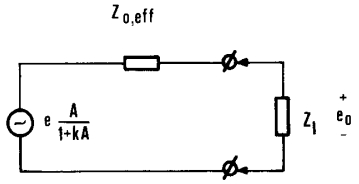


Fig. 74-34. – Equivalent circuit on the output side in Figure 74-33.

which is smaller than Z_o. Hence, this negative feedback both raises input impedance and lowers output impedance.

The above-discussed feedback, where the voltage subtracted is proportional to output voltage, is called *voltage feedback*. If the subtracted voltage is made proportional to output current, it is called *current feedback*. Applying negative current feedback, one finds that this raises the input impedance but also raises the output impedance, the former also being dependent on the load at the output of the amplifier.

In the above, a *voltage* proportional to an output quantity was used as a feedback signal; if, instead, a *current* is used as a feedback signal, things are somewhat different. Note that the commonly used terms "voltage" and "current" feedback refer to which of these at the amplifier's output governs the feedback. Now we will study the effect of different feedback quantities at the input; we will discuss "feedback voltage" and "feedback current." The feedback given first above is the effect of a "voltage-feedback-voltage." Below, we will see a "voltage-feedback-current" (a "current-feedback-current" will have an effect on input impedance that depends on the load on the amplifier, like the "current-feedback-voltage" mentioned above, and therefore will not be discussed here).

If a feedback *current* proportional to the output *voltage* is added to the current coming from the source at the input, as in Figure 74-35, we find:

$$Z_{\text{in, eff}} = \frac{Z_{\text{in}}}{1 - k_1 A Z_{\text{in}}} \tag{74-91}$$

with a constant k_1. Hence, in this case, *positive* feedback increases input impedance. Obviously, this means an inherent risk for instability. The stability criterion actually depends on Z_s. In principle, infinite $Z_{\text{in, eff}}$ is possible in equation (74-91) if $k_1 A Z_{\text{in}} = 1$ in the passband of the arrangement, but, in practice, this may call for unrealizable stability requirements on the characteristic outside the passband, especially as this is *positive* feedback, although stability requirements may be fulfilled inside the passband. Further, in the case when a "feedback voltage" is used, the increase of input impedance may be limited by practical stability requirements on roll-offs, although *negative* feedback there makes things easier.

Noise and Impedance Relations

Similar to the derivation of an optimal relation between source and load impedances for maximal power transfer, leading to impedance matching (equation [74-83]), one can find an optimal impedance relation at the input for the best noise figure of the particular source-amplifier arrangement; this is again related to "power economy," due to the definition of noise figure. In biomedical applications, one usually must measure a voltage from a *given* source where one has little or no possibility of altering the source impedance. For "voltage economy," we found that the input impedance of the amplifier should be as high as possible; a similar rule exists for noise "voltage economy."

An amplifier may be characterized, in terms of "noisiness," by referring the noise to the input (as mentioned above). The general representation then involves both a noise voltage source in series and a noise current source in parallel with the input (in which case also a "cross power spectrum" between the two sources is specified). For simplicity, we will show only the case of a pure voltage source, as sketched in Figure 74-36, where e_{ns} is the noise voltage generated in the source, which we assume cannot be influenced, and e_n is the amplifier noise referred to the input. For this case, we find

$$e_o = A \frac{Z_{\text{in}}}{Z_{\text{in}} + Z_s}\left(e + e_{ns} + e_n \frac{Z_s}{Z_{\text{in}}}\right) \tag{74-92}$$

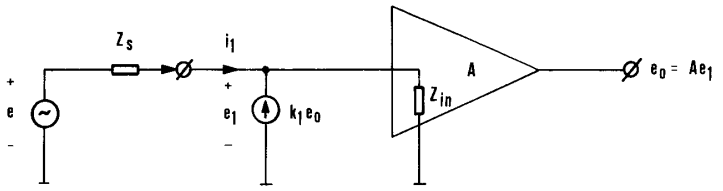
As Z_s and e_{ns} are given, we can improve the signal-to-noise ratio at the amplifier output only by choosing an amplifier having as small a value as possible of e_n/Z_{in}, unless we apply feedback. Feedback may lead to an improvement, raising Z_{in}, but then noise sources in the feedback network also add to total noise at the input. Therefore, feedback may instead occasionally make things worse with respect to noise. This must be judged in each individual case. As a general rule, it is advisable first to choose an amplifier with the highest Z_{in}/e_n in the important frequency range, and then to see if feedback may bring about further improvement.

In the general case with both a voltage and a current source in the representation at the output, we also find that noise sources should be as small and input impedance as high as possible. If i_n is the noise current source, we should have $(i_n + e_n/Z_{\text{in}})$ as small as possible.

"Bootstrapping" of Cable Capacitance

When a long cable connects source and amplifier, it usually is shielded to reduce pickup of power-line hum or other interfering signals. (Magnetically induced interference still

Fig. 74-35. – Influence of a feedback current.



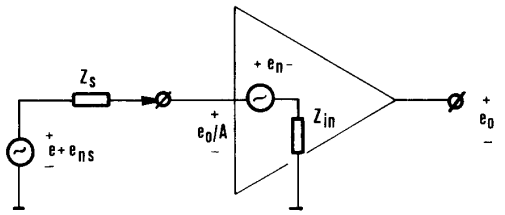


Fig. 74-36. – Influence of source impedance on signal-to-noise ratio; e_{ns} is the source noise and e_n is the amplifier noise, referred to the input.

may occur, as well as that due to movement of the cable in the static magnetic field of the earth, or the electrostatic charging and discharging in the insulation caused by friction between it and the central lead wire.) Such a *coaxial cable* has a relatively high capacitance between lead and shield; if the latter is grounded, it may cause deterioration of the input impedance, as seen from the source. One useful method to considerably reduce this effect is to connect the shield to a signal source so that it receives a reproduction of the amplifier's input voltage. This signal can be obtained from the output of the first stage if it has unity amplification. For small signals, it is better to have high amplification in the first stage, then reduce the signal level again to the input level (e.g., with a voltage divider) and then apply it to the shield. The output impedance of this connection presented to the shield must be much lower than the amplifier's source and input impedances. A second, grounded, shield may be used outside or around the ungrounded one.

This arrangement often is called a "hot," "driven" or "pumped" shield and reduces the effect of the cable capacitance by keeping the voltage over it at zero, thus eliminating capacitive currents. This is a special type of feedback, often called "bootstrapping." The effect is less useful if the shield does not enclose the lead.

A number of errors in the type setting have been corrected.

At many locations, the typesetter has mistaken a 1 for an l (lower case L) in the formulae.

These locations are too many to search for and correct (in which process some would have been overlooked, anyway). Therefore, where for example l/f should read 1/f, and so on, the reader will surely notice it...

The article was published as Chapter 74 in *Medical Engineering*, edited by Charles D. Ray on pp. 974 – 1003, Year Book Medical Publishers, Inc., Chicago, 1974.