4

BASEBAND COMMUNICATION

Baseband communication refers to signal transmission *without modulation*; the name stems from the fact that baseband transmission does not involve frequency translation of the message spectrum that characterizes modulation. And although the majority of communication systems are modulation systems, baseband transmission deserves our study. For one reason, baseband connecting links are part of most modulation systems; for another, many of the concepts and parameters of baseband communication carry over directly to modulation. But perhaps most important, the performance characteristics of baseband transmission serve as useful standards when comparing the various types of modulation.

This chapter, therefore, is devoted to baseband communication. It begins with an investigation of the two fundamental limitations of electrical signaling mentioned in Chap. 1, namely, noise and bandwidth. The results are then applied to three distinct classes of baseband transmission: analog, pulse, and digital. The coverage ranges from elementary but significant design calculations to optional discussion of sophisticated optimization techniques.

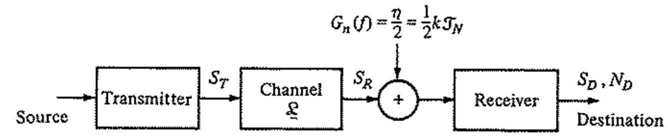


FIGURE 4.1
Parameters of a communication system.

4.1 SIGNALS AND NOISE

We observed in Chap. 3 that noise becomes a significant factor in electrical communication when the received signal is very feeble and therefore of the same order of magnitude as the ever-present thermal noise. In turn, the very small received signal level is due primarily to the large amount of power loss that characterizes long-haul transmission. Hence, our discussion of signals and noise starts with a consideration of transmission loss. Figure 4.1 puts this topic in context and locates the various system parameters with which we shall deal. These parameters are defined as follows:

S_T = signal power at the transmitter output
$\mathcal{L}$ = transmission power loss of the channel
S_R = signal power at the receiver input
$\mathcal{T}_N$ = noise temperature referred to the receiver input
η = noise density (assumed constant) at the receiver input
S_D = signal power at the destination
N_D = noise power at the destination

This set of notation is used throughout the rest of the text and the reader should carefully study the definitions.

Transmission Loss

When we say that an amplifier has *power gain* $\mathcal{G}$, we mean that the input and output powers are related by $P_{\text{out}} = \mathcal{G}P_{\text{in}}$, so

$$\mathcal{G} \triangleq \frac{P_{\text{out}}}{P_{\text{in}}} \tag{1a}$$

Frequently gain is expressed in decibels† as

$$\mathcal{G}_{\text{dB}} \triangleq 10 \log_{10} \mathcal{G} = 10 \log_{10} \frac{P_{\text{out}}}{P_{\text{in}}} \tag{1b}$$

† See Table E for decibel conversions and manipulations.

Hence, if $P_{out} = 100P_{in}$, $\mathscr{G} = 100$ and $\mathscr{G}_{dB} = 20$ dB; if the power gain is unity, then $\mathscr{G}_{dB} = 0$ dB. But almost all transmission channels have $P_{out} < P_{in}$ and $\mathscr{G} < 1$. Thus we define the *power loss*

$$\mathscr{L} \triangleq \frac{1}{\mathscr{G}} = \frac{P_{in}}{P_{out}} \tag{2a}$$

$$\mathscr{L}_{dB} = -\mathscr{G}_{dB} = 10 \log_{10} \frac{P_{in}}{P_{out}} \tag{2b}$$

also called the *attenuation*.

Although we have defined $\mathscr{G}$ and $\mathscr{L}$ as power ratios, they can be expressed as ratios of mean-square voltage or current, e.g.,

$$\mathscr{G} = \frac{\overline{v_{out}^2}}{\overline{v_{in}^2}} \tag{3}$$

where the bar stands for either time average or ensemble average, as appropriate. Equation (3) is consistent with our use of normalized power, and we will not get involved in the distinctions between available power gain, transducer power gain, etc.

Illustrating Eq. (2), the input-output relations for transmission lines, coaxial cables, and waveguides are all of the form $10^{-0.868\alpha l} P_{in}$

$$P_{out} = e^{-2\alpha l} P_{in} = 10^{-8.68\alpha l} P_{in} \tag{4}$$

where α is the attenuation coefficient and l is the path length. Therefore the loss in decibels is $0.868\alpha l$

$$\mathscr{L}_{dB} = 8.68\alpha l \tag{5}$$

Table 4.1 lists some representative values. Actually, Eqs. (4) and (5) strictly hold only for sinusoidal signals and α depends on the frequency, as the table implies. For the time being we ignore the frequency dependence since the point here is the potentially

Table 4.1 TYPICAL VALUES OF TRANSMISSION LOSS

Transmission medium	Frequency	Loss, dB/km
Open-wire pair (0.3 cm diameter)	1 kHz	0.05
Twisted-wire pair (16 gauge)	10 kHz	2
	100 kHz	3
	300 kHz	6
Coaxial cable (1 cm diameter)	100 kHz	1
	1 MHz	2
	3 MHz	4
Coaxial cable (15 cm diameter)	100 MHz	1.5
Rectangular waveguide (5 × 2.5 cm)	10 GHz	5
Helical waveguide (5 cm diameter)	100 GHz	1.5

large values of loss. Underscoring that point, a 20-km run of 16-gauge twisted pair at 100 kHz has $\mathscr{L}_{dB} = 3 \times 20 = 60$ dB or $\mathscr{L} = 10^6$, which means that $P_{out} = P_{in}/\mathscr{L} = 10^{-6} P_{in}$! Also observe that the decibel loss is directly proportional to l; so, continuing our example, a 40-km run has $\mathscr{L}_{dB} = 2 \times 60 = 120$ dB and $P_{out} = 10^{-12} P_{in}$!!

In view of the above numbers, it is not surprising that transmission by radio propagation is often preferred for large distances. Although radio systems involve modulation, to be covered in later chapters, it seems appropriate here to mention the difference between cable and radio transmission insofar as path loss is concerned.

Specifically, the power ratio on a line-of-sight radio path is

$$\frac{P_{out}}{P_{in}} = \mathscr{G}_{TA} \mathscr{G}_{RA} \left(\frac{\lambda}{4\pi l}\right)^2 \tag{6}$$

where $\mathscr{G}_{TA}$ and $\mathscr{G}_{RA}$ are the power gains of the transmitting and receiving antennas and λ is the wavelength of the carrier. Thus

$$\mathscr{L}_{dB} = 22 + 10 \log_{10} \left(\frac{l}{\lambda}\right)^2 - (\mathscr{G}_{TA_{dB}} + \mathscr{G}_{RA_{dB}}) \tag{7}$$

so a 20-km path with $\lambda = 1$ meter and $\mathscr{G}_{TA} = \mathscr{G}_{RA} = 16$ dB has $\mathscr{L}_{dB} = 22 + 86 - 32 = 76$ dB. But note that Eq. (7) is *not* proportional to l; in fact, doubling l increases the loss by only 6 dB (an assertion the reader should confirm for himself). Therefore, a 40-km path with the above parameters has $\mathscr{L}_{dB} = 76 + 6 = 82$ dB. Incidentally, 40 km is just about the upper limit for line-of-sight paths over flat terrain unless rather high antenna towers are used.

Returning to Fig. 4.1, we can now express S_R in terms of S_T and $\mathscr{L}$ simply as

$$S_R = \frac{S_T}{\mathscr{L}} \tag{8}$$

Then, if the power gain of the receiver is $\mathscr{G}_R$,

$$S_D = \mathscr{G}_R S_R = \frac{\mathscr{G}_R}{\mathscr{L}} S_T \tag{9a}$$

Equations such as Eq. (9*a*) may be written in the form

$$S_{D_{dBW}} = \mathscr{G}_{R_{dB}} - \mathscr{L}_{dB} + S_{T_{dBW}} \tag{9b}$$

where dBW stands for decibels above 1 watt, e.g., 20 dBW = 100 W. Hence, multiplication and division are replaced by addition and subtraction of decibel values, which is why they are so much used by communication engineers. But the extra subscripts in Eq. (9*b*), etc., are a nuisance and will be omitted hereafter; it should be clear from the context when dB values are implied.

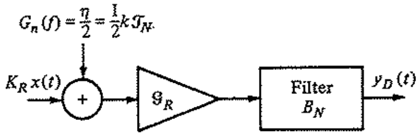


FIGURE 4.2
Model of receiver with additive white noise.

Additive Noise

Having gotten the signal to the receiver, let us look at the effects of contaminating noise. Figure 4.2 extracts the pertinent portion of Fig. 4.1. The waveform at the receiver input is $K_R x(t)$, where $x(t)$ is the information-bearing signal, to which is added white noise having $G_n(f) = \eta/2 = k\mathcal{T}_N/2$. The noise temperature $\mathcal{T}_N$ represents *all* the noise in the system, referred to the receiver input.† Numerical values may range from around 60°K in a carefully engineered low-noise system to several thousand degrees.

For analysis purposes, the receiver has been separated into two parts: an amplifier with power gain $\mathcal{G}_R$ (and voltage gain $\sqrt{\mathcal{G}_R}$) followed by a filter with noise equivalent bandwidth B_N and $\mathcal{G} = 1$. The function of the filter is to pass $x(t)$ but reject as much noise as possible, namely, those noise components outside the signal's frequency range—the *out-of-band* noise, in other words.

Assuming no nonlinearities in the receiver, the total output is

$$y_D(t) = \sqrt{\mathcal{G}_R}\, K_R x(t) + n_D(t) \qquad (10)$$

where $n_D(t)$ is the output noise. It is not unreasonable to further assume that $x(t)$ and $n_D(t)$ are *statistically independent* and that $\bar{n}_D = 0$; under these conditions, $x(t)$ and $n_D(t)$ are *incoherent* and their crosscorrelation is zero, i.e., $R_{xn_D}(\tau) = R_{n_Dx}(\tau) = 0$. Therefore, using Eq. (10), Sect. 3.5,

$$R_{y_D}(\tau) = \mathcal{G}_R K_R^2 R_x(\tau) + R_{n_D}(\tau) \qquad (11)$$

and

$$\overline{y_D^2} = \underbrace{\mathcal{G}_R K_R^2 \overline{x^2}}_{S_D} + \underbrace{\overline{n_D^2}}_{N_D} \qquad (12)$$

where we have identified the two components of $\overline{y_D^2}$ as the output signal and noise powers, respectively. Specifically, since $S_R = K_R^2\overline{x^2}$,

$$S_D = \mathcal{G}_R K_R^2 \overline{x^2} = \mathcal{G}_R S_R \qquad (13)$$

† See Appendix B for the method.

and, from Eq. (14), Sect. 3.6,

$$N_D = \overline{n_D^2} = \mathcal{G}_R \eta B_N \qquad (14)$$

in which one can insert $\eta = k\mathcal{T}_N$ if desired. If the noise is not white, then

$$N_D = \mathcal{G}_R \int_{-\infty}^{\infty} |H_R(f)|^2 G_n(f)\, df$$

where $H_R(f)$ describes the filter.

Signal-to-Noise Ratios

Whenever we have an expression like Eq. (12), it is meaningful to speak of the ratio of signal power to noise power—or *signal-to-noise ratio*, for short. In the case at hand, we define the signal-to-noise ratio at the destination

$$\left(\frac{S}{N}\right)_D \triangleq \frac{S_D}{N_D} = \frac{K_R^2\overline{x^2}}{\eta B_N} = \frac{S_R}{\eta B_N} \qquad (15)$$

Note that the receiver gain $\mathcal{G}_R$ has canceled out, being common to both terms. Such cancellation will always occur, and the only function of $\mathcal{G}_R$ is to produce the desired signal level at the output.

On the other hand, any gains or losses that enter the picture *before* the noise has been added will definitely affect $(S/N)_D$. As an important case in point, recall that $S_R = S_T/\mathcal{L}$, so

$$\left(\frac{S}{N}\right)_D = \frac{S_T}{\mathcal{L}\eta B_N} \qquad (16)$$

and we see that the signal-to-noise ratio is inversely proportional to the transmission loss.

One final comment here regarding the assumptions behind these results: if the noise is not additive, or if the receiver has nonlinearities, or if $x(t)$ and $n_D(t)$ are not incoherent, then $R_{y_D}(\tau)$ will include *crosscorrelations* $R_{xn_D}(\tau)$, etc. Therefore, $\overline{y_D^2}$ will have crossproduct or signal-times-noise terms so the definition of the signal-to-noise ratio is ambiguous and not particularly meaningful.

Repeater Systems

If the path loss on a given system yields an unsatisfactorily low value for $(S/N)_D$ and the parameters S_T, η, and B_N are fixed, there still remains an alternative, namely, the introduction of one or more amplifier-filter units between source and destination. These units are called *repeaters*.

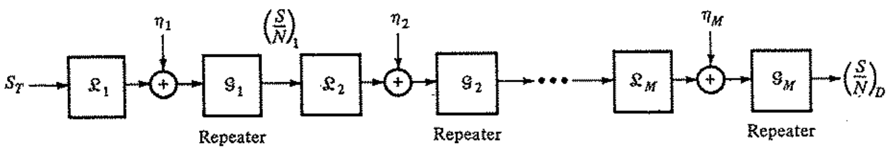


FIGURE 4.3
Repeater system.

A repeater system is block-diagramed in Fig. 4.3, where the filters are combined with the amplifiers. The last repeater and the receiver are one and the same. Typical repeater spacings are as small as 2 km for certain cable systems, and run up to 40 km for microwave radio relay systems. Hence, with the notable exception of communication satellite links, it takes from 100 to 2,000 repeaters to span the North American continent.

Inspecting Fig. 4.3 shows that the S/N at the output of the first repeater is

$$\left(\frac{S}{N}\right)_1 = \frac{S_T}{\mathscr{L}_1 \eta_1 B_N}$$

Usually repeater systems are designed with identical units and just enough gain to overcome the loss, i.e., $\mathscr{G}_1/\mathscr{L}_1 = \mathscr{G}_2/\mathscr{L}_2 = \cdots = 1$. Furthermore, $\eta_1 = \eta_2 = \cdots = \eta_M$ when the units are identical and the amplifier noise dominates other noise sources. Under these conditions† $S_D = S_T$, $N_D = M\mathscr{G}_1\eta_1 B_N = M\mathscr{L}_1\eta_1 B_N$, and

$$\left(\frac{S}{N}\right)_D = \frac{S_T}{M\mathscr{L}_1\eta_1 B_N} = \frac{1}{M}\left(\frac{S}{N}\right)_1 \qquad (17)$$

Therefore, the destination S/N equals $1/M$ times the S/N for one link or hop.

It might appear from Eq. (17) that little has been gained by using repeaters. Bear in mind, however, that $\mathscr{L}_1$ is the loss on just *one* hop; if repeaters are not used, the total loss is $\mathscr{L} = \mathscr{L}_1\mathscr{L}_2 \cdots \mathscr{L}_M = \mathscr{L}_1{}^M$. Hence, $\mathscr{L}_1 = \mathscr{L}^{1/M}$, which represents a substantial loss reduction, as the following exercise will reveal.

EXERCISE 4.1 A signal is to be transmitted 40 km using a transmission line whose loss is 3 dB/km; the receiver has $\mathscr{T}_N = 10\mathscr{T}_0$ and $B_N = 5$ kHz. Calculate ηB_N in decibels above one watt (dBW) and find the value of S_T (in watts) required to get $(S/N)_D = 50$ dB. Repeat the second calculation when there is a repeater at the halfway point. *Ans.*: -157 dBW, 20 W, 40 μW.

† Appendix B gives the analysis of the general case.

4.2 SIGNAL DISTORTION IN TRANSMISSION

Besides noise, the other fundamental limitation of electrical communication is bandwidth, the finite bandwidth of any real systems that leads to signal distortion. But *distortionless transmission* does not necessarily imply that the output is identical to the input. Certain differences can be tolerated and not classified as distortion. Our purpose here is to formalize the meaning of distortionless transmission and the requirements for it. With this background, the various types of distortion can be defined and their effects investigated. The emphasis will be on those aspects pertinent to communication systems.

Stated crudely, for distortionless transmission the output should "look like" the input. More precisely, given an input signal $x(t)$, we say that the output is undistorted if it differs from the input only by a multiplying constant and a finite time delay. Analytically, we have distortionless transmission if

$$y(t) = Kx(t - t_d) \qquad (1)$$

where K and t_d are constants.

The properties of a distortionless network are easily found by examining the output spectrum

$$Y(f) = \mathscr{F}[y(t)] = Ke^{-j\omega t_d}X(f)$$

Now by definition of transfer function, $Y(f) = H(f)X(f)$, so

$$H(f) = Ke^{-j\omega t_d} \qquad (2a)$$

In words, a network giving distortionless transmission must have *constant amplitude response* and negative *linear phase shift*, that is,

$$|H(f)| = K \qquad \arg[H(f)] = -2\pi t_d f \pm m180° \qquad (2b)$$

We have added the $\pm m180°$ term to account for the constant being positive or negative. Zero phase is allowable since it implies zero time delay. One more qualification can be added to Eq. (2): these conditions are required only over those frequencies for which the input signal has nonzero spectrum. Thus, if $x(t)$ is bandlimited in W, Eq. (2) need be satisfied only for $|f| < W$.

In practice distortionless transmission is a stringent condition which, at best, can be only approximately satisfied. Thus, an inescapable fact of signal transmission is that distortion will occur, though it can be minimized by proper design. We should therefore be concerned with the *degree* of distortion, measured in some quantitative fashion. Unfortunately, quantitative measures prove to be rather unwieldy and impractical for engineering purposes. As an alternate approach, distortion has been classified as to type, and each type considered separately. But before discussing the various types, it must be emphasized that *distortion is distortion*; a severely distorted output will differ significantly from the input, regardless of the specific cause.

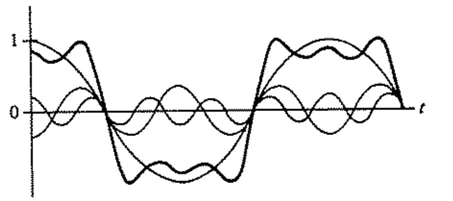


FIGURE 4.4
Test signal $x(t) = \cos \omega_0 t - \frac{1}{3} \cos 3\omega_0 t + \frac{1}{5} \cos 5\omega_0 t$.

The three major classifications of distortion are:

1 Amplitude distortion: $|H(f)| \neq K$

2 Phase (delay) distortion: $\arg [H(f)] \neq -2\pi t_d f \pm m180°$

3 Nonlinear distortion

The first two cases are categorized as *linear* distortion. In the third case the system includes nonlinear elements, and its transfer function is not defined. We now examine these individually.

Amplitude Distortion

Amplitude distortion is easily described in the frequency domain; it means simply that the output frequency components are not in correct proportion. Since this is caused by $|H(f)|$ not being constant with frequency, amplitude distortion is sometimes called *frequency distortion.*

The most common forms of amplitude distortion are excess attenuation or enhancement of extreme high or low frequencies in the signal spectrum. Less common, but equally bothersome, is disproportionate response to a band of frequencies within the spectrum. While the frequency-domain description is easy, the effects in the time domain are far less obvious, save for very simple signals. For illustration, a suitably simple test signal is $x(t) = \cos \omega_0 t - \frac{1}{3} \cos 3\omega_0 t + \frac{1}{5} \cos 5\omega_0 t$, Fig. 4.4, a rough approximation to a square wave. If the low-frequency or high-frequency component is attenuated by one-half, the resulting outputs are as shown in Fig. 4.5. As expected, loss of the high-frequency term reduces the "sharpness" of the waveform.

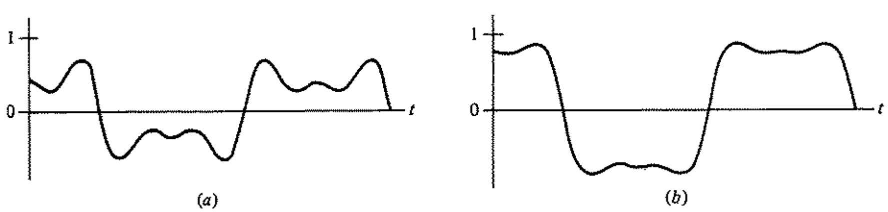
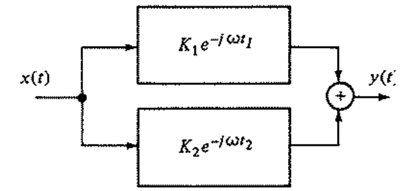


FIGURE 4.5
Test signal with amplitude distortion. (*a*) Low frequency attenuated; (*b*) high frequency attenuated.

FIGURE 4.6
Block diagram of radio system with two propagation paths.



Beyond qualitative observations, there is little more that can be said about amplitude distortion without experimental study of specific signal types. Results of such studies are usually couched in terms of required *frequency response*, i.e., the range of frequencies over which $|H(f)|$ must be constant to within a certain tolerance (say ± 1 dB) so that the amplitude distortion is sufficiently small.

EXERCISE 4.2 Sometimes radio systems suffer from *multipath distortion* caused by two (or more) propagation paths between transmitter and receiver. As a simple example, suppose that the received signal is

$$y(t) = K_1 x(t - t_1) + K_2 x(t - t_2) \tag{3a}$$

Show that Fig. 4.6 is the equivalent block diagram and that, if $(K_2/K_1)^2 \ll 1$,

$$|H(f)| \approx K_1 \left[1 + \frac{K_2}{K_1} \cos 2\pi f (t_2 - t_1)\right] \tag{3b}$$

Hence, a "weak" reflection yields *ripples* in the amplitude ratio.

Phase Shift and Delay Distortion

A linear phase shift yields a constant time delay for all frequency components in the signal. This, coupled with constant amplitude response, yields an undistorted output. If the phase shift is not linear, the various frequency components suffer different amounts of time delay, and the resulting distortion is termed *phase* or *delay distortion.*

For an arbitrary phase shift, the time delay is a function of frequency, call it $t_d(f)$, and can be found by writing $\arg [H(f)] = -2\pi f t_d(f)$, so

$$t_d(f) = -\frac{\arg [H(f)]}{2\pi f} \tag{4}$$

which is independent of frequency only if $\arg [H(f)]$ is linear with frequency.

A common area of confusion is *constant time delay* versus *constant phase shift.* The former is desirable and is required for distortionless transmission. The latter, in general, causes distortion. Suppose a system has the constant phase shift θ. Then

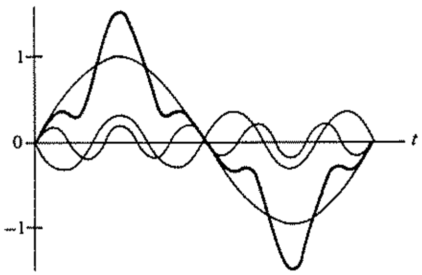


FIGURE 4.7
Test signal with constant phase shift of −90°.

each signal frequency component will be delayed by $\theta/2\pi$ *cycles* of its own frequency; this is the meaning of constant phase shift. But the time delays will be different, the frequency components will be scrambled in time, and distortion will result. However, the constant phase shifts $\theta = 0$ and $\pm m180°$ are acceptable.

That constant phase shift does give distortion is simply illustrated by returning to the test signal of Fig. 4.4 and shifting each component by one-fourth cycle, $\theta = -90°$. Whereas the input was roughly a square wave, the output will look like a triangular wave, Fig. 4.7. With an arbitrary nonlinear phase shift, the deterioration of waveshape can be even more severe.

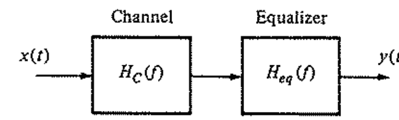
One should also note from Fig. 4.7 that the *peak* excursions of the phase-shifted signal are substantially greater (by about 50 percent) than those of the input test signal. This is not due to amplitude response, since the output amplitudes of the three frequency components are, in fact, unchanged; rather, it is because the components of the distorted signal all attain maximum or minimum values at the same time, which was not true of the input. Conversely, had we started with Fig. 4.7 as the test signal, a constant phase shift of +90° would yield Fig. 4.4 for the output waveform. Thus we see that *delay distortion alone* can result in an increase or decrease of peak values as well as other waveshape alterations.

Clearly, delay distortion can be critical in pulse transmission, and much labor is spent *equalizing* transmission delay for digital data systems and the like. On the other hand, the human ear is curiously insensitive to delay distortion; the waveforms of Figs. 4.4 and 4.7 would sound just about the same when driving a loudspeaker. Thus, delay distortion is seldom of concern in voice and music transmission.

Equalization

Linear distortion—i.e., amplitude and delay distortion—is theoretically curable through the use of *equalization* networks. Figure 4.8 shows an equalizer $H_{eq}(f)$ in cascade with a distorting channel $H_C(f)$. Since the overall transfer function is $H(f) =$

FIGURE 4.8
Channel with equalizer for linear distortion.



$H_C(f)H_{eq}(f)$ the final output will be distortionless if $H_C(f)H_{eq}(f) = Ke^{-j\omega t_d}$, where K and t_d are more or less arbitrary constants. Therefore, we require that

$$H_{eq}(f) = \frac{Ke^{-j\omega t_d}}{H_C(f)} \tag{5}$$

wherever $X(f) \neq 0$.

Rare is the case when an equalizer can be designed to satisfy Eq. (5) exactly—which is why we say that equalization is a *theoretical* cure. But excellent approximations often are possible so that linear distortion can be reduced to a tolerable level. Probably the oldest equalization technique is the use of *loading coils* on twisted-pair telephone lines.† These coils are lumped inductors placed in shunt across the line every kilometer or so, giving the improved amplitude ratio typically illustrated in Fig. 4.9. Other lumped-element circuits have been designed for specific equalization tasks.

More recently, the *tapped-delay-line equalizer* or *transversal filter* has emerged as a convenient and flexible device. To illustrate the principle, Fig. 4.10 shows a delay line with total time delay 2Δ having taps at each end and the middle. The tap outputs are passed through adjustable gains, c_{-1}, c_0, and c_1, and summed to form the final output. Thus

$$y(t) = c_{-1}x(t) + c_0 x(t-\Delta) + c_1 x(t-2\Delta) \tag{6a}$$

and

$$\begin{aligned} H_{eq}(f) &= c_{-1} + c_0 e^{-j\omega\Delta} + c_1 e^{-j\omega 2\Delta} \\ &= (c_{-1}e^{+j\omega\Delta} + c_0 + c_1 e^{-j\omega\Delta})e^{-j\omega\Delta} \end{aligned} \tag{6b}$$

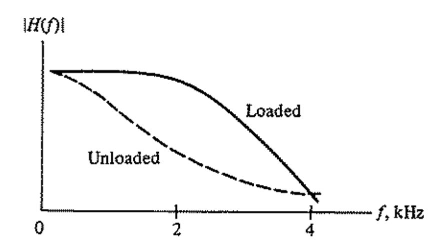


FIGURE 4.9
Amplitude ratio of a typical telephone line with and without loading coils for equalization.

† Everitt and Anner (1956, chap. 8) gives the theory of loading.

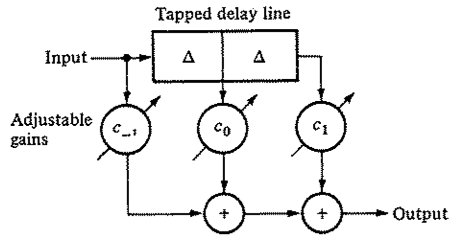


FIGURE 4.10
Tapped-delay-line equalizer (transversal filter) with three taps.

Clearly, this is a convenient arrangement since the tap gains are more readily changed than lumped elements. To demonstrate the flexibility, suppose $c_{-1} = c_1 < c_0/2$; then

$$|H_{eq}(f)| = c_0 + 2c_1 \cos \omega\Delta \qquad \arg [H_{eq}(f)] = -\omega\Delta$$

On the other hand, if $c_{-1} = -c_1$ and $|c_1| \ll c_0$, then

$$|H_{eq}(f)| \approx c_0 \qquad \arg [H_{eq}(f)] \approx -\omega\Delta - \frac{2c_1}{c_0} \sin \omega\Delta$$

Therefore, depending on the tap gains, we can equalize amplitude ripples or phase ripples or both.

Generalizing Eq. (6*b*) to the case of a $2M\Delta$ delay line with $2M + 1$ taps,

$$H_{eq}(f) = \left(\sum_{m=-M}^{M} c_m e^{-j\omega m\Delta} \right) e^{-j\omega M\Delta} \tag{7}$$

which has the form of an *exponential Fourier series* with frequency periodicity $1/\Delta$. Therefore, given a channel $H_C(f)$ to be equalized over $|f| < W$, one can approximate the right-hand side of Eq. (5) by a Fourier series with frequency periodicity $1/\Delta \geq W$ (thereby determining Δ), estimate the number of significant terms (which determines M), and match the tap gains to the series coefficients. However, this high-powered theoretical method may not be needed in simple cases such as the following example.

Example 4.1

Suppose we wish to equalize the multipath distortion described in Exercise 4.2, where

$$H_C(f) = K_1 \left[1 + \frac{K_2}{K_1} e^{-j\omega(t_2 - t_1)} \right] e^{-j\omega t_1}$$

with $(K_2/K_1)^2 \ll 1$ and $t_2 > t_1$. Applying Eq. (5), the equalizer should have

$$H_{eq}(f) = \frac{K}{K_1} \frac{e^{-j\omega(t_d - t_1)}}{1 + (K_2/K_1)e^{-j\omega(t_2 - t_1)}}$$

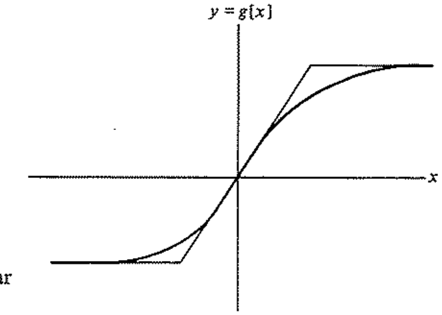


FIGURE 4.11
Transfer characteristic of a nonlinear device.

Taking $K = K_1$ and $t_d = t_1$ and expanding the denominator as a three-term binomial series yields

$$H_{eq}(f) \approx 1 - \frac{K_2}{K_1} e^{-j\omega(t_2 - t_1)} + \left(\frac{K_2}{K_1}\right)^2 e^{-j\omega(t_2 - t_1)} e^{-j\omega(t_2 - t_1)}$$

$$\approx \left[e^{+j\omega(t_2 - t_1)} - \frac{K_2}{K_1} + \left(\frac{K_2}{K_1}\right)^2 e^{-j\omega(t_2 - t_1)} \right] e^{-j\omega(t_2 - t_1)}$$

Comparing this with Eq. (6*b*) reveals that a three-tap transversal filter will do the job if $c_{-1} = 1$, $c_0 = -(K_2/K_1)$, $c_1 = (K_2/K_1)^2$, and $\Delta = t_2 - t_1$. ////

Nonlinear Distortion

A system having nonlinear elements cannot be described by a transfer function. Instead, the instantaneous values of input and output are related by a curve or function $y(t) = g[x(t)]$, commonly called the *transfer characteristic.* Figure 4.11 is a representative transfer characteristic; the flattening out of the output for large input excursions is the familiar saturation-and-cutoff effect of transistor amplifiers. We shall consider only *memoryless* devices, for which the transfer characteristic is a complete description.

Under small-signal input conditions, it may be possible to linearize the transfer characteristic in a piecewise fashion, as shown by the thin lines in the figure. The more general approach is a polynomial approximation to the curve, of the form

$$y(t) = a_1 x(t) + a_2 x^2(t) + a_3 x^3(t) + \cdots \tag{8a}$$

It is the higher powers of $x(t)$ in this equation that give rise to the nonlinear distortion.

Even though we have no transfer function, the output spectrum can be found, at least in a formal way, by transforming Eq. (8*a*). Specifically, invoking the convolution theorem,

$$Y(f) = a_1 X(f) + a_2 X * X(f) + a_3 X * X * X(f) + \cdots \tag{8b}$$

Now if $x(t)$ is bandlimited in W, the output of a linear network will contain no frequencies beyond $|f| < W$. But in the nonlinear case, we see that the output includes $X * X(f)$, which is bandlimited in $2W$, $X * X * X(f)$, which is bandlimited in $3W$, etc. The nonlinearities have therefore created output frequency components that were not present in the input. Furthermore, since $X * X(f)$ may contain components for $|f| < W$, this portion of the spectrum overlaps that of $X(f)$. Using filtering techniques, the added components at $|f| > W$ can be removed, but there is no convenient way to get rid of the added components at $|f| < W$. These, in fact, constitute the nonlinear distortion.

A quantitative measure of nonlinear distortion is provided by taking a simple cosine wave, $x(t) = \cos \omega_0 t$, as the input. Inserting in Eq. (8*a*) and expanding yields

$$y(t) = \left(\frac{a_2}{2} + \frac{3a_4}{8} + \cdots\right) + \left(a_1 + \frac{3a_3}{4} + \cdots\right) \cos \omega_0 t + \left(\frac{a_2}{2} + \frac{a_4}{4} + \cdots\right) \cos 2\omega_0 t + \cdots$$

Therefore, the nonlinear distortion appears as *harmonics* of the input wave. The amount of second-harmonic distortion is the ratio of the amplitude of this term to that of the fundamental, or in percent:

$$\text{Second-harmonic distortion} = \frac{a_2/2 + a_4/4 + \cdots}{a_1 + 3a_3/4 + \cdots} \times 100\%$$

Higher-order harmonics are treated similarly. However, their effect is usually much less, and many can be removed entirely by filtering.

If the input is a sum of two cosine waves, say $\cos \omega_1 t + \cos \omega_2 t$, the output will include all the harmonics of f_1 and f_2, plus crossproduct terms which yield $f_2 - f_1, f_2 + f_1, f_2 - 2f_1$, etc. These sum and difference frequencies are designated as *intermodulation distortion*. Generalizing the intermodulation effect, if $x(t) = x_1(t) + x_2(t)$, then $y(t)$ contains the *cross product* $x_1(t)x_2(t)$ (and higher-order products, which we ignore here). In the frequency domain $x_1(t)x_2(t)$ becomes $X_1 * X_2(f)$; and even though $X_1(f)$ and $X_2(f)$ may be separated in frequency, $X_1 * X_2(f)$ can overlap both of them, producing one form of *cross talk*. This aspect of nonlinear distortion is of particular concern in telephone transmission systems. On the other hand the crossproduct term is the desired result when nonlinear devices are used for modulation purposes.

Companding

Although nonlinear distortion has no perfect cure, it too can be minimized by careful design. The basic idea is to make sure that the signal does not exceed the linear operating range of the channel's transfer characteristic. Ironically, one strategy along this

FIGURE 4.12
Companding system.

line involves *two nonlinear* devices, a *compressor* and an *expander*, arranged per Fig. 4.12.

A compressor is a device having greater amplification at low signal levels than at high signal levels. Since amplification is the derivative of the transfer characteristic with respect to the input, a typical compressor characteristic $g_{comp}[x(t)]$ would be as shown in Fig. 4.13. Note that a compressor compresses the range of the output signal. Therefore, if the compressed range falls within the linear range of the channel, the signal at the channel output is proportional to $g_{comp}[x(t)]$ which is distorted by the compressor but not the channel. Ideally, then, the expander should have a characteristic that perfectly complements the compressor—i.e., less amplification at low signal levels, etc. Thus, the final output is proportional to $g_{exp}\{g_{comp}[x(t)]\} = x(t)$, as desired.

The joint use of compressing and expanding is called *companding* (surprise?) and is of particular value in telephone systems. Besides combating nonlinear distortion, companding tends to compensate for the signal-level difference between loud and soft talkers. Indeed, the latter is the key advantage of companding compared to the simpler technique of linearly attenuating the signal at the input (to keep it in the linear range of the channel) and linearly amplifying it at the output.

4.3 ANALOG TRANSMISSION

Getting it all together, as it were, this section applies the results of the previous two sections to the case of analog transmission. By *analog transmission* we mean those systems in which information-bearing waveforms are to be reproduced at the destination without employing digital coding techniques.

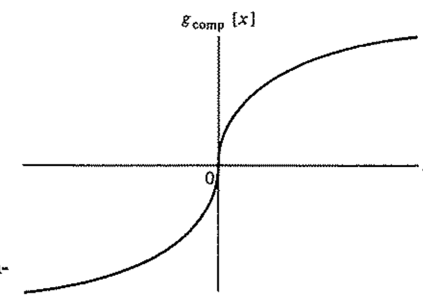


FIGURE 4.13
Typical transfer characteristic of a compressor.

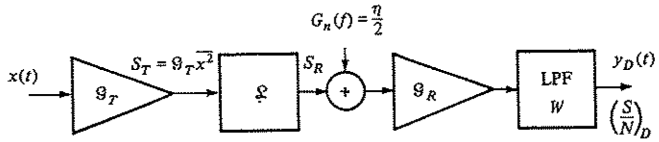


FIGURE 4.14
Analog transmission system.

Our analyses will be in terms of an arbitrary waveform or *message* designated by $x(t)$. Better yet, $x(t)$ represents the ensemble of probable messages from a given source. Though such messages are not strictly bandlimited, it is safe to assume that there exists some upper frequency—call it W—above which the spectral content is negligible and unnecessary for conveying the information in question. Thus, we define

$$W = \text{analog message bandwidth}$$

in the sense that

$$G_x(f) \approx 0 \qquad \text{for } |f| > W \tag{1}$$

We further assume ergodicity so that $\langle x^2(t)\rangle$ and $\overline{x^2}$ are interchangeable.

Signal-to-Noise Ratio

Figure 4.14 amplifies Fig. 4.1 for the situation in question. Specifically, the transmitter becomes simply an amplifier with power gain $\mathcal{G}_T$, so $S_T = \mathcal{G}_T\overline{x^2}$, and the receiver filter is a nearly ideal LPF with bandwidth W, so $B_N \approx W$. The other parameters are the same as defined in Sect. 4.1.

If the total transmission delay is t_d and there is no distortion over $|f| \leq W$, the output signal is

$$y_D(t) = \left(\frac{\mathcal{G}_T\mathcal{G}_R}{\mathcal{L}}\right)^{1/2} x(t - t_d) + n_D(t) \tag{2}$$

so the destination signal and noise powers are

$$S_D = \frac{\mathcal{G}_T\mathcal{G}_R}{\mathcal{L}} \mathrm{E}[x^2(t - t_d)] = \frac{\mathcal{G}_T\mathcal{G}_R}{\mathcal{L}} \overline{x^2}$$

$$N_D = \overline{n_D{}^2} = \mathcal{G}_R \eta W$$

Therefore

$$\left(\frac{S}{N}\right)_D = \frac{\mathcal{G}_T\overline{x^2}}{\mathcal{L}\eta W} = \frac{S_T}{\mathcal{L}\eta W} \tag{3}$$

or, since $S_T/\mathcal{L} = S_R$,

$$\left(\frac{S}{N}\right)_D = \frac{S_R}{\eta W} \tag{4}$$

Table 4.2 lists representative values of $(S/N)_D$ for selected analog signals, along with the frequency range. The upper limit of the frequency range is the nominal value of W. The lower limit also has design significance since, because of transformers and coupling capacitors, most analog transmission systems do not respond all the way down to DC.

Equation (4) expresses $(S/N)_D$ in terms of some very basic system parameters, namely, the signal power and noise density at the receiver input and the message bandwidth. This combination of terms will occur over and over again, particularly when we compare various system types, so we give it a symbol of its own by defining

$$\gamma \triangleq \frac{S_R}{\eta W} \tag{5}$$

In the present context γ equals $(S/N)_D$ for analog baseband transmission. We can also interpret the denominator ηW as the *noise power in the message bandwidth*, even though N_D differs from ηW by $\mathcal{G}_R$. (Recall that this gain factor cancels out in signal-to-noise ratios.)

Because Eq. (4) presupposes distortionless transmission conditions, additive white noise, and a nearly ideal filter, it is more accurate to say that

$$\left(\frac{S}{N}\right)_D \leq \gamma \tag{6}$$

In other words, γ generally is an *upper bound* for analog baseband performance that may or may not be achieved in an actual system. For instance, the noise bandwidth of a practical LPF will be somewhat greater than the message bandwidth, giving $(S/N)_D = S_R/\eta B_N < \gamma$. Similarly, nonlinearities that cause the output to include signal-

Table 4.2 TYPICAL TRANSMISSION REQUIREMENTS FOR SELECTED ANALOG SIGNALS

Signal type	Frequency range	Signal-to-noise ratio, dB
Barely intelligible voice	500 Hz–2 kHz	5–10
Telephone-quality voice	200 Hz–3.2 kHz	25–35
AM broadcast-quality audio	100 Hz–5 kHz	40–50
High-fidelity audio	20 Hz–20 kHz	55–65
Television video	60 Hz–4.2 MHz	45–55

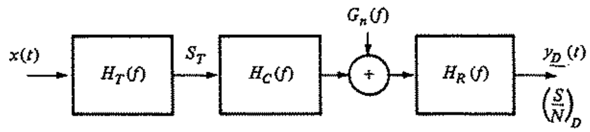


FIGURE 4.15

times-noise terms also reduce the effective S/N. Companding, on the other hand, may yield a net improvement.† The effects of linear distortion and nonwhite noise are examined jointly below.

Optimum Terminal Filters ★

When the noise is nonwhite and/or the channel requires considerable equalization, the rather simple-minded approach taken above should be replaced by a more sophisticated technique in which specially designed filters are incorporated at both terminals, the transmitter and receiver. Figure 4.15 is the system diagram, with the power gains $\mathscr{G}_T$, $1/\mathscr{L}$, and $\mathscr{G}_R$ this time absorbed in the frequency-response functions $H_T(f)$, $H_C(f)$, and $H_R(f)$.

As far as distortionless transmission is concerned, any pair of terminal filters will do, providing

$$H_T(f)H_R(f) = \frac{Ke^{-j\omega t_d}}{H_C(f)} \qquad |f| < W \tag{7}$$

Hence, if $H_R(f)$ is chosen to minimize the output noise and $H_T(f)H_R(f)$ satisfies Eq. (7), we have *optimized* the terminal filters in the sense that $(S/N)_D$ is maximum and the output signal is undistorted. But the optimization has a subtle constraint; namely, the transmitted power S_T must be kept within reasonable bounds. Accordingly, we seek to minimize $S_T N_D/S_D$ rather than N_D alone.

Assuming Eq. (7) holds, the total output signal is

$$y_D(t) = Kx(t - t_d) + n_D(t)$$

and

$$S_D = K^2\overline{x^2} \tag{8}$$

$$N_D = \int_{-\infty}^{\infty} |H_R(f)|^2 G_n(f)\, df \tag{9}$$

† See Bennett (1970, chap. 3).

At the transmitting end we take $G_x(f)$ as the message spectral density so

$$S_T = \int_{-\infty}^{\infty} |H_T(f)|^2 G_x(f)\, df = \int_{-\infty}^{\infty} \frac{K^2 G_x(f)}{|H_C(f)H_R(f)|^2}\, df \tag{10}$$

where Eq. (7) has been used to eliminate $|H_T(f)|^2$. Therefore, the quantity to minimize is

$$\frac{S_T N_D}{S_D} = \frac{1}{\overline{x^2}}\left[\int_{-\infty}^{\infty} \frac{G_x}{|H_C H_R|^2}\, df \int_{-\infty}^{\infty} |H_R|^2 G_n\, df\right] \tag{11}$$

it being understood that all terms are functions of f except $\overline{x^2}$, which is a constant. Note that the only function in Eq. (11) under the designer's control is the receiving filter $H_R(f)$.

Normally, optimization problems require the methods of variational calculus. But this particular problem (and a few others that follow) can be solved by adroit application of *Schwarz's inequality*, Eq. (6), Sect. 2.6, one form of which is

$$\left|\int_{-\infty}^{\infty} VW^*\, df\right|^2 \le \int_{-\infty}^{\infty} |V|^2\, df \int_{-\infty}^{\infty} |W|^2\, df \tag{12a}$$

where V and W are arbitrary functions of f and the equality holds when

$$V(f) = \text{constant} \times W(f) \tag{12b}$$

Now, except for the constant $\overline{x^2}$, Eq. (11) has the same form as the right-hand side of Eq. (12*a*) with

$$V = \frac{G_x^{1/2}}{|H_C H_R|} \qquad W = |H_R| G_n^{1/2}$$

both of which are real and nonnegative. Thus $S_T N_D/S_D$ is minimized when

$$|H_R(f)|_{\text{opt}}^2 = \frac{G_x^{1/2}(f)}{|H_C(f)|\, G_n^{1/2}(f)} \tag{13a}$$

as follows by taking $V(f) = W(f)$, the proportionality constant being immaterial (why?). Equation (13*a*) is the optimum receiving filter, and the corresponding optimum transmitting filter is, from Eq. (7),

$$|H_T(f)|_{\text{opt}}^2 = \frac{K^2 G_n^{1/2}(f)}{|H_C(f)|\, G_x^{1/2}(f)} \tag{13b}$$

with K^2 being determined from the desired value for S_T.

Interpreting these equations, we see that $H_R(f)$ deemphasizes those frequencies where the noise density is large and the signal density is small—a very sensible thing to do—while $|H_T(f)|$ does just the reverse. The phase shift of $H_T(f)$ and $H_R(f)$ does not appear here since we are dealing with spectral densities, but the overall phase must satisfy Eq. (7) for distortionless transmission.

Finally, with optimum filtering the destination S/N is

$$\left(\frac{S}{N}\right)_{D_{\max}} = \frac{S_T}{(S_T N_D/S_D)_{\min}} = \frac{S_T \overline{x^2}}{\left|\int_{-\infty}^{\infty} \frac{G_x^{1/2}(f) G_n^{1/2}(f)}{|H_C(f)|} df\right|^2} \tag{14}$$

as follows from Eqs. (11) and (12*a*).

Aside from the question of synthesizing the filters, the major obstacle preventing complete optimization in practice is the assumption implied by Eq. (13) that $G_x(f)$ is known in detail. Usually, the communication engineer does know the general characteristics of the message—or, rather, the class or ensemble of possible messages—but not the complete details. For instance, if the messages are known to be band-limited in W, one might then assume that the spectral density is flat over $|f| < W$, i.e.,

$$G_x(f) = \frac{\overline{x^2}}{2W} \Pi\left(\frac{f}{2W}\right) \tag{15}$$

there being no reason to believe that $G_x(f)$ is larger or smaller at any particular frequency. Proceeding on this assumption would yield a good design but not necessarily optimum.

Actually, Eq. (15) is the underlying assumption of our previous approach where we took the noise to be white and said that the receiving filter serves only to eliminate the out-of-band noise. Clarifying this point, let $G_n(f) = \eta/2$ and let the channel be distortionless so $|H_C(f)|^2 = 1/\mathscr{L}$. Then, inserting in Eq. (13) shows that the terminal filters become ideal LPFs while the denominator of Eq. (14) is

$$\left|\int_{-W}^{W} (\mathscr{L}\overline{x^2}\eta/4W)^{1/2}\, df\right|^2 = \mathscr{L}\overline{x^2}\eta W$$

so $(S/N)_{D_{\max}} = S_T/\mathscr{L}\eta W = S_R/\eta W = \gamma$.

EXERCISE 4.3 Consider a system having $|H_c(f)|^2 = 1/\mathscr{L}$ and $G_n(f) = (\eta/2)(1 + a^2 f^2)^2$. Taking $G_x(f)$ per Eq. (15), find $(S/N)_{D_{\max}}$ in terms of S_R and compare with $(S/N)_D$ when the receiving filter is an ideal LPF with $B = W$. *Ans.*: $(S/N)_{D_{\max}} = S_R/\eta W[1 + (a^2W^2/3)]^2$, $(S/N)_D = S_R/\eta W[1 + (2a^2W^2/3) + (a^4W^4/5)]$.

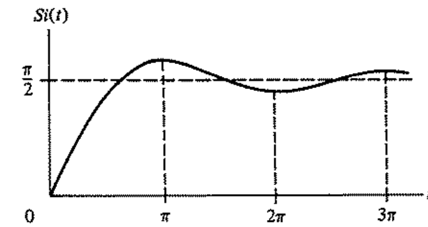


FIGURE 4.16
The sine integral, $\mathrm{Si}(t) = \int_0^t (\sin\lambda)/\lambda\, d\lambda$.

4.4 PULSE TRANSMISSION

Pulse transmission differs from analog transmission in that one is not concerned with the faithful reproduction of a message waveform but rather with detecting the presence of a pulse, resolving two or more closely spaced pulses, and measuring the amplitude or time position. Telegraph and radar systems are examples. This section examines the effects of limited bandwidth and additive noise on pulse transmission at baseband.

Bandwidth Requirements

Short pulses have large spectral widths, as we have seen time and again. Reversing this observation, it can be said that given a system of fixed bandwidth, there is a lower limit on the duration of pulses at the output, i.e., a *minimum output pulse duration.* Consequently, the maximum number of distinct output pulses that can be resolved per unit time is limited by the system bandwidth.

To put the matter on a quantitative footing, let a rectangular pulse $x(t) = A\Pi(t/\tau)$ be the input to an ideal or nearly ideal LPF with bandwidth B, unit gain, and zero time delay, so $H(f) = \Pi(f/2B)$. Since the input spectrum $X(f) = A\tau \operatorname{sinc} f\tau$ has even symmetry, the inverse Fourier transform for the output $y(t) = \mathscr{F}^{-1}[H(f)X(f)]$ simplifies to

$$y(t) = 2\int_0^B A\tau \frac{\sin \pi f\tau}{\pi f\tau} \cos 2\pi ft\, df$$
$$= \frac{A}{\pi}\left[\int_0^B \frac{\sin \pi f(2t+\tau)}{f} df - \int_0^B \frac{\sin \pi f(2t-\tau)}{f} df\right]$$

which is still a nonelementary integral requiring series evaluation. Fortunately, the result can be expressed in terms of the tabulated *sine integral*

$$\mathrm{Si}(t) \triangleq \int_0^t \frac{\sin\lambda}{\lambda} d\lambda \tag{1}$$

plotted in Fig. 4.16. Changing integration variables finally leads to

$$y(t) = \frac{A}{\pi}\{\mathrm{Si}[\pi B(2t+\tau)] - \mathrm{Si}[\pi B(2t-\tau)]\} \tag{2}$$

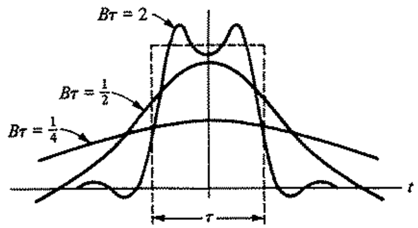


FIGURE 4.17
Pulse response of an ideal LPF.

which is shown in Fig. 4.17 for three values of the product $B\tau$. Notice the precursors caused by the ideal filter.

Despite the rather involved mathematics, the conclusions drawn from Fig. 4.17 are quite simple. We have said that the spectral width of a rectangular pulse is about $1/\tau$. For $B \gg 1/\tau$, the output signal is essentially undistorted; whereas for $B \ll 1/\tau$, the output pulse is stretched and has a duration that depends more on the filter bandwidth than on the input signal. As a rough but highly useful rule of thumb one can say that the minimum output pulse duration and bandwidth are related by

$$\tau_{\min} \geq \frac{1}{2B} \tag{3}$$

providing the input pulse has $\tau \leq \tau_{\min}$. Going somewhat further, we can also say that the maximum number of resolved output pulses per unit time is about $1/\tau_{\min} = 2B$. This is achieved using input pulses of duration less than $1/2B$ and spaced in time by $1/2B$. Figure 4.18, showing the input and output signals for two pulses spaced by τ, supports this assertion.

Beside pulse detection and resolution, one may be concerned with the question of pulse location or position measured with respect to some reference time. Usually, position measurements are based on the leading edge of the pulse and for that purpose rectangular pulses would be desired since the edge has a unique position. But realizable pulse shapes rise more gradually toward their peak value, causing the position of the leading edge to be ambiguous and its measurement less certain. The conventional

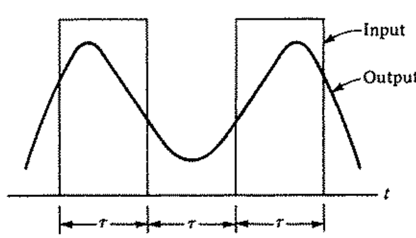


FIGURE 4.18
Pulse resolution of an ideal LPF, $B = 1/2\tau$.

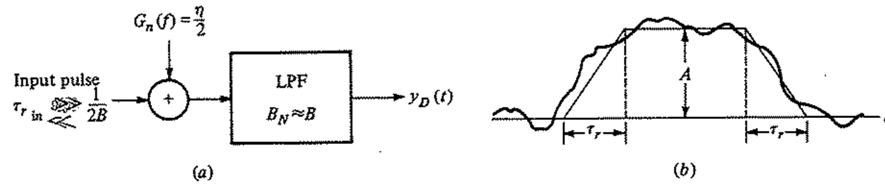


FIGURE 4.19
Pulse measurements in additive noise. (*a*) Block diagram; (*b*) filtered pulse with noise.

rule about uncertainty is stated in terms of the *rise time*, defined as the interval required for the pulse to go from zero to full amplitude or from 10 to 90 percent of full amplitude. We then say that the uncertainty of the pulse position measurement approximately equals the rise time τ_r. Referring back to Fig. 4.17 it is seen that the rise time of a filtered pulse is proportional to the bandwidth. Therefore, as another rule of thumb we have

$$\tau_{r_{\min}} \geq \frac{1}{2B} \tag{4}$$

When the input pulses have $\tau_r \leq \tau_{r_{\min}}$, the output pulses will have rise times no less than $1/2B$ and the minimum location uncertainty is about $1/2B$. Alternately, if the input rise time is greater than $1/2B$, the output rise time will be approximately the same as the input.

Granted that Eqs. (3) and (4) are rough guidelines based on the case of a rectangular input to an ideal LPF, they are nonetheless useful in general. Studies of other pulse shapes and other lowpass filters show that these inequalities are appropriate, taking B as the 3-dB bandwidth.

Pulse Measurements in Additive Noise

Let a pulse (not necessarily rectangular) be contaminated by additive white noise and passed through an LPF whose noise equivalent bandwidth and 3-dB bandwidth are approximately equal, Fig. 4.19*a*. If the rise time of the input pulse is small compared to $1/2B$, the output pulse shape can be approximated as a *trapezoid* plus noise $n_D(t)$, Fig. 4.19*b*. Thus, measurements of both the pulse amplitude and position will be in error owing to the noise.

At the peak of the output pulse, $y_D(t) = A + n_D(t)$, so we can define the normalized mean-square amplitude error as

$$\epsilon_A^2 \triangleq \frac{\overline{n_D^2}}{A^2} = \frac{\eta B}{A^2} \tag{5a}$$

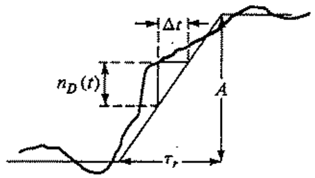


FIGURE 4.20
Expanded view of noise perturbation.

But the energy of the output pulse is $E \approx A^2\tau$ and $\tau \geq 1/2B$; hence

$$\epsilon_A{}^2 \geq \frac{\eta}{2E} \tag{5b}$$

which gives a lower bound on the error. This lower bound holds when $\tau = 1/2B$.

Position measurements usually are accomplished by noting the time at which the output pulse exceeds some fixed level, say $A/2$. Then, as seen in the expanded view of Fig. 4.20, the time position error is $\Delta t = (\tau_r/A)n_D(t)$—from the similar triangles—and the mean-square error normalized by τ^2 is

$$\epsilon_t{}^2 \triangleq \frac{\overline{\Delta t^2}}{\tau^2} = \overline{n_D{}^2}\,\frac{\tau_r{}^2}{A^2\tau^2} = \eta B\,\frac{\tau_r{}^2}{A^2\tau^2} \tag{6a}$$

Finally, inserting $\tau_r \geq 1/2B$ and $E = A^2\tau$ yields the lower bound

$$\epsilon_t{}^2 \geq \frac{\eta}{4BE\tau} \tag{6b}$$

In contrast to Eq. (5*b*), this lower bound is achieved when $\tau_r = 1/2B$ and $\tau \gg \tau_r$. Thus, unlike amplitude measurement, the largest possible bandwidth should be used to minimize position measurement errors.

By now, the reader may have recognized that our analysis of pulse transmission is largely intuitive and heuristic. More refined investigations of specific cases are possible, notably the case of optimum pulse detection discussed below. However, the admittedly crude results above still are valuable guidelines for the design of pulse transmission systems, and the reciprocals of Eqs. (5) and (6) parallel the concept of signal-to-noise ratio in analog transmission.

Optimum Pulse Detection—Matched Filters ★

Similar to the optimum terminal filters for analog transmission, there exists an optimum receiving filter for detecting a pulse of known shape $x(t)$ contaminated by additive noise with known spectral density $G_n(f)$. Such filters are termed *matched filters*, used extensively in radar and data transmission systems.

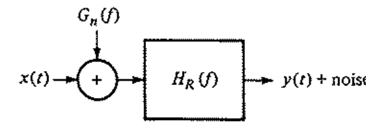


FIGURE 4.21

Consider the situation in Fig. 4.21, where the output pulse shape is unimportant but one desires to maximize its amplitude at some arbitrary time, say t_0, and minimize the output noise. In absence of noise the peak output signal at $t = t_0$ is

$$y(t_0) = \mathscr{F}^{-1}[H(f)X(f)]|_{t=t_0}$$
$$= \int_{-\infty}^{\infty} H(f)X(f)e^{+j\omega t_0}\,df \tag{7}$$

where $X(f) = \mathscr{F}[x(t)]$. (We use the Fourier transforms rather than the spectral density since the pulse is a known energy signal.) The output noise power is

$$N = \int_{-\infty}^{\infty} |H(f)|^2 G_n(f)\,df \tag{8}$$

and the quantity to be maximized is

$$\frac{|y(t_0)|^2}{N} = \frac{|\int_{-\infty}^{\infty} HXe^{j\omega t_0}\,df|^2}{\int_{-\infty}^{\infty} |H|^2 G_n\,df} \tag{9}$$

where $H(f)$ is the only function at our disposal.

To determine $H_{opt}(f)$, we again draw upon Schwarz's inequality, Eq. (12), Sect. 4.3, this time in the form

$$\frac{|\int_{-\infty}^{\infty} VW^*\,df|^2}{\int_{-\infty}^{\infty} |V|^2\,df} \leq \int_{-\infty}^{\infty} |W|^2\,df$$

whose left-hand side is the same as Eq. (9) with

$$V = HG_n{}^{1/2} \qquad W^* = \frac{HXe^{j\omega t_0}}{V} = \frac{Xe^{j\omega t_0}}{G_n{}^{1/2}}$$

Since the inequality becomes an equality when $V(f) = KW(f)$, the ratio in Eq. (9) will be maximized if

$$H_{opt}(f) = K\,\frac{X^*(f)e^{-j\omega t_0}}{G_n(f)} \tag{10}$$

where K is an arbitrary constant, as is t_0. Therefore,

$$\left[\frac{|y(t_0)|^2}{N}\right]_{max} = \int_{-\infty}^{\infty} \frac{|X(f)|^2}{G_n(f)}\,df \tag{11}$$

if the filter is optimized.

Observe from Eq. (10) that $H_{opt}(f)$ emphasizes those frequencies where $|X(f)|/G_n(f)$ is large, and vice versa, similar to the optimum receiving filter for analog transmission, Eq. (13*a*), Sect. 4.3. Unfortunately, $H_{opt}(f)$ often turns out to be physically *unrealizable*† because the corresponding impulse response is nonzero for $t < 0$. The following exercise relates Eq. (11) to our previous studies and shows why $H_{opt}(f)$ is called a matched filter.

EXERCISE 4.4 For the case of white noise $G_n(f) = \eta/2$, show that

$$\left[\frac{|y(t_0)|^2}{N}\right]_{max} = \frac{2E_x}{\eta} \qquad (12a)$$

where E_x is the energy in $x(t)$. Also show that

$$h_{opt}(t) = \mathscr{F}^{-1}[H_{opt}(f)] = Kx(t_0 - t) \qquad (12b)$$

so the impulse response has the same shape as the input pulse reversed in time and shifted by t_0. (*Hint*: Use the fact that $v(-t) \leftrightarrow V^*(f)$ when $v(t)$ is real.)

4.5 DIGITAL TRANSMISSION

We conclude this chapter with a brief examination of digital transmission at baseband. Fundamentally, a digital message is nothing more than an ordered sequence of *symbols* drawn from an *alphabet* of finite size μ. (For instance, a binary source has $\mu = 2$ and the alphabet symbols are the digits **0** and **1**.) The objective of a digital communication system is to transmit the message in a prescribed amount of time with a minimum number of errors. Thus, *signaling rate* and *error probability* play the same role in digital transmission that bandwidth and signal-to-noise ratio play in analog transmission. Moverover, there is a close relationship between signaling rate and bandwidth, and between error probability and signal-to-noise ratio.

Waveforms and Signaling Rate

One normally thinks of a digital signal as being a string of discrete-amplitude rectangular pulses. And, in fact, that is often the way it comes from the data source. By way of illustration, Fig. 4.22*a* shows the binary message **10110100** as it might appear at the output of a digital computer. This waveform, a simple on-off sequence, is said to be *unipolar*, because it has only one polarity, and *synchronous*, because all

† Thomas (1969, chap. 5) investigates the optimization problem with a realizability constraint.

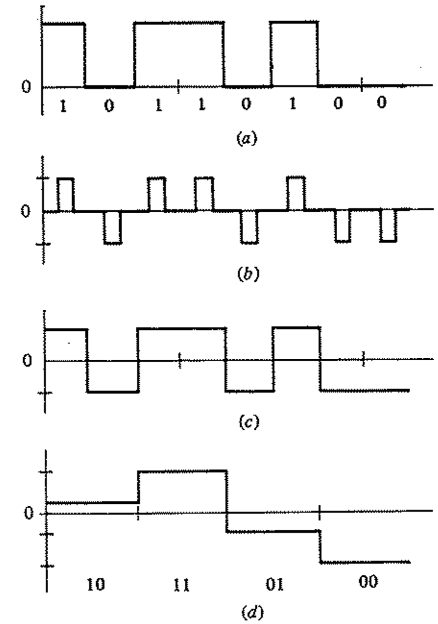


FIGURE 4.22
Digital waveforms. (*a*) Unipolar synchronous; (*b*) polar return-to-zero; (*c*) polar synchronous; (*d*) polar synchronous quaternary.

pulses have equal duration and there is no separation between them. Unipolar signals contain a nonzero DC component that is difficult to transmit, carries no information, and is a waste of power. Similarly, synchronous signals require timing coordination at transmitter and receiver, which means design complications. The *polar* (two-polarity) *return-to-zero* signal of Fig. 4.22*b* gets around both of these problems, but the "spaces" making the signal self-clocking are a waste of transmission time. If efficiency is a dominant consideration, the *polar synchronous* signal of Fig. 4.22*c* would be preferable. Illustrating a multilevel case, Fig. 4.22*d* is a quaternary ($\mu = 4$) signal derived by grouping the binary digits in blocks of two.

Regardless of the specific details, the channel input signal is an analog representation of the digital message that can generally be described as a pulse train of the form†

$$x(t) = \sum_k a_k p\left(t - \frac{k}{r}\right) \qquad (1)$$

where a_k is the amplitude level representing the kth message digit, $p(t)$ is the basic pulse shape with peak value $p(0) = 1$, the pulse-to-pulse spacing is $1/r$, and r is the

† The index k indicates time sequence, and its limits, omitted in Eq. (1), depend on when the message starts and stops.

signaling rate. For example, the polar synchronous signal of Fig. 4.22c has $a_k = \pm a$ and $p(t) = \Pi(t/\tau)$, where the pulse duration is $\tau = 1/r$. A return-to-zero signal would have $\tau < 1/r$.

If the channel is linear and distortionless over all frequencies—i.e., has infinite bandwidth—then $p(t)$ suffers no degradation in transmission, and an arbitrarily large signaling rate can be achieved by using very short pulses. But a real channel has finite bandwidth and less than ideal frequency response, causing the pulses to spread out and overlap. The engineer must therefore shape the output signal so as to *minimize intersymbol interference* due to overlapping and, at the same time, *maximize signaling rate*, objectives that are mutually contradictory.

This problem has been studied since the earliest days of telegraphy, but it was Harry Nyquist (1924, 1928) who first stated the bandwidth–signaling rate relationship:

> Given an ideal lowpass channel of bandwidth B, it is possible to send independent symbols at a rate $r \le 2B$ symbols per second without intersymbol interference. It is not possible to send independent symbols at $r > 2B$.

Note that $r \le 2B$ agrees with the pulse-resolution rule $\tau_{\min} \ge 1/2B$ of Sect. 4.4 since $\tau \le 1/r$.

It is an easy matter to prove the second part of the relationship, for suppose we try to signal at $2(B + \epsilon)$ symbols per second, ϵ being positive but arbitrarily small. One possible message sequence consists of two symbols alternating indefinitely, **01010101** ..., for example. The resulting channel waveform is *periodic* with period $1/(B + \epsilon)$ and contains only the fundamental frequency $f_0 = B + \epsilon$ plus its harmonics. Since no frequency greater than B is passed by the channel, the channel output will be zero—aside from a possible but useless DC component.

Signaling at the maximum rate $r = 2B$ requires a very special pulse shape, namely, the *sinc pulse*

$$p(t) = \operatorname{sinc} rt \qquad (2)$$

which is *bandlimited* in $B = r/2$ and therefore suffers no distortion when transmitted over the channel. Of course $p(t)$ is not timelimited, but it does have *periodic zero crossings*, i.e.,

$$p\left(\frac{m}{r}\right) = \operatorname{sinc} m = \begin{cases} 1 & m = 0 \\ 0 & m \ne 0 \end{cases}$$

Thus, if we form the signal

$$x(t) = \sum_k a_k \operatorname{sinc} r\left(t - \frac{k}{r}\right) = \sum_k a_k \operatorname{sinc}(rt - k) \qquad (3a)$$

then at any time $t = m/r$,

$$x\left(\frac{m}{r}\right) = \sum_k a_k \operatorname{sinc}(m - k) = a_m \qquad (3b)$$

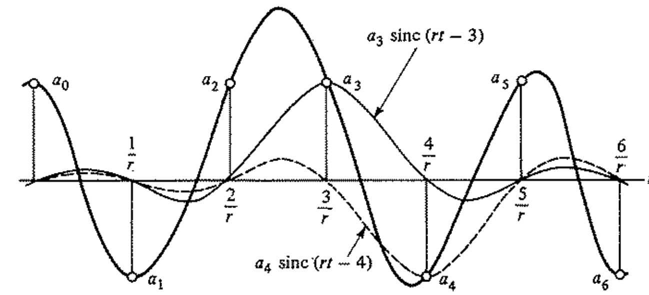


FIGURE 4.23
The digital waveform $x(t) = \sum_k a_k \operatorname{sinc}(rt - k)$.

as illustrated in Fig. 4.23. In other words, because of the zero crossings, the overlapping pulses do not result in intersymbol interference if we periodically sample $x(t)$ at the rate $r = 2B$.

As the reader may have inferred, timing information between transmitter and receiver is required here—i.e., the signaling must be synchronized at exactly $r = 2B$. Furthermore, this approach only works with an ideal lowpass channel. The general question of pulse shaping is deferred to Chap. 10; for the time being we will take $x(t)$ as in Eq. (3), which happens to be the most efficient choice if the bandwidth is limited. On the other hand, if the available bandwidth is large compared to r, then rectangular pulses would be the most expedient choice.

EXERCISE 4.5 Show that synchronous signaling at $r = B$ is possible if $p(t) = \operatorname{sinc}^2 rt$.

Noise and Errors

Figure 4.24 shows the basic elements of a digital baseband receiver. The received signal $K_R x(t)$ is contaminated by additive noise, and a nearly ideal LPF removes the out-of-band noise to yield

$$y(t) = K_R x(t) + n(t)$$

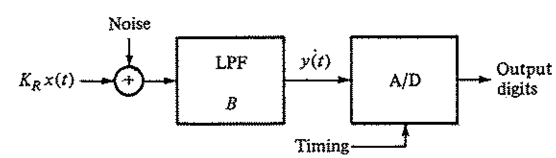


FIGURE 4.24
Baseband digital receiver.

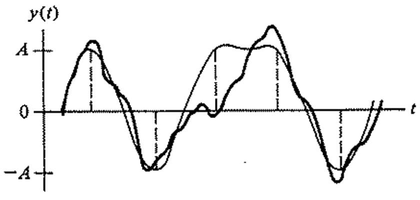


FIGURE 4.25
Polar binary waveform plus noise.

where the D subscripts have been dropped for convenience. The analog signal $y(t)$ is then operated on by an analog-to-digital (A/D) converter whose function is to recover or *regenerate* the digital message. Synchronization is supplied to the A/D converter so it can sample $y(t)$ at the optimum times $t = m/r$ when there is no intersymbol interference, i.e., $y(m/r) = K_R a_m + n(m/r)$.

To begin with a simple case, consider a received polar binary signal having $K_R a_k = \pm A$ representing the binary digits **1** and **0**. A typical signal-plus-noise waveform $y(t)$ as it might appear at the input to the A/D converter is illustrated in Fig. 4.25, assuming $\bar{n} = 0$. A direct conversion technique is to decide, at the appropriate times, whether $y(m/r)$ is closer to $+A$ (a **1** presumably intended) or closer to $-A$ (a **0** presumably intended). Intuitively, the logical *decision rule* becomes: choose **1** if $y(m/r) > 0$, choose **0** if $y(m/r) < 0$, and flip a coin if $y(m/r) = 0$. (This last event, being rare, will receive no further attention.) The converter can therefore take the form of a synchronized decision circuit whose crossover or *threshold level* is set at zero, and conversion errors occur whenever the noise causes $y(m/r)$ to be on the wrong side of the threshold at the decision time.

Thus, insofar as error probabilities are concerned, we have *two* random variables†

$$y_1 = A + n \qquad y_0 = -A + n \tag{4}$$

corresponding to the intended digits **1** and **0**. Then, if a **1** was intended, the *conditional probability* of conversion error is

$$P_{e_1} \triangleq P(\text{error}\,|\,\mathbf{1}\text{ sent}) = P(y_1 < 0) = P(A + n < 0) \tag{5a}$$

and similarly

$$P_{e_0} \triangleq P(\text{error}\,|\,\mathbf{0}\text{ sent}) = P(y_0 > 0) = P(-A + n > 0) \tag{5b}$$

Hence, the *net error probability* becomes

$$P_e = P_1 P_{e_1} + P_0 P_{e_0} \tag{6}$$

where P_1 and P_0 are the digit probabilities at the source, not necessarily equal but usually so. In any case $P_1 + P_0 = 1$, since one or the other must be transmitted.

† Here, for simplicity, we use lowercase letters to symbolize random variables.

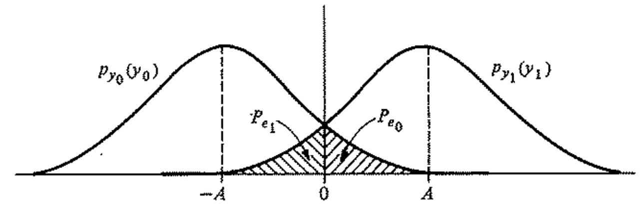


FIGURE 4.26
Signal-plus-noise PDFs for polar binary signaling.

At this point we assume that $n(t)$ is a zero-mean *gaussian* process with variance σ^2, so its probability density function is

$$p_n(n) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-n^2/2\sigma^2} \tag{7}$$

This assumption is not unreasonable for linear baseband systems since most electrical noise is gaussian and gaussian functions are invariate under linear operations. It follows from Eq. (4) that y_1 and y_0 are also gaussian with the same variance σ^2 but with mean values

$$\bar{y}_1 = +A \qquad \bar{y}_0 = -A$$

Figure 4.26 conveniently summarizes the situation by showing the two PDFs, $p_{y_1}(y_1) = p_n(y_1 - A)$ and $p_{y_0}(y_0) = p_n(y_0 + A)$, from which we will calculate P_{e_1} and P_{e_0}.

Recalling the area interpretation of probability density functions,

$$\begin{aligned} P_{e_0} = P(y_0 > 0) &= \int_0^\infty p_{y_0}(y_0)\, dy_0 \\ &= \frac{1}{\sqrt{2\pi}\,\sigma} \int_0^\infty e^{-(y_0 + A)^2/2\sigma^2}\, dy_0 \\ &= \frac{1}{\sqrt{2\pi}\,\sigma} \int_{A/\sigma}^\infty e^{-\lambda^2/2}\, d\lambda \end{aligned}$$

where the change of variable $\lambda = (y_0 + A)/\sigma$ has been made. This puts P_{e_0} in the same form as the function $Q(\kappa)$, Eq. (8), Sect. 3.4, with $\kappa = A/\sigma$. Furthermore, noting the symmetry of Fig. 4.26, $P(y_1 < 0) = P(y_0 > 0)$ so

$$P_{e_1} = P_{e_0} = Q\left(\frac{A}{\sigma}\right) \tag{8a}$$

Therefore, from Eq. (6), the net error probability is

$$P_e = (P_1 + P_0) Q\left(\frac{A}{\sigma}\right) = Q\left(\frac{A}{\sigma}\right) \tag{8b}$$

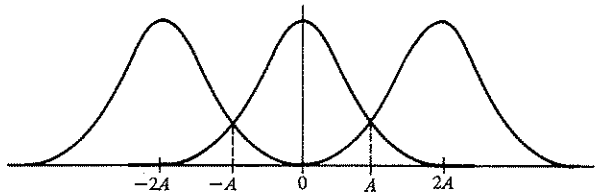


FIGURE 4.27
Signal-plus-noise PDFs for polar trinary signaling.

which is independent of the digit probabilities because $P_{e_1} = P_{e_0}$. Several points in this analysis deserve further comment:

1 Taking the threshold level at zero yields *equal error probabilities* for each digit, $P_{e_1} = P_{e_0}$.
2 This is the *optimum* threshold level if the digits are equiprobable ($P_1 = P_0$) since any other choice would increase P_{e_1} more than it decreases P_{e_0}, or vice versa.
3 Binary signals in gaussian noise have a unique property apparent in Fig. 4.26. If the intended amplitude is $+A$, superimposed positive noise excursions have no detrimental effect; and similarly for $-A$ with negative noise excursions. Hence, because $n(t)$ is equally likely to be positive or negative, the converter will be correct at least half the time, regardless of the noise. However, one should bear in mind that a binary message with 50 percent errors is 100 percent worthless.

4 In view of Eq. (8*b*), Fig. 3.5 or Table D may be used directly as a plot of P_e versus A/σ and it is evident that P_e decreases dramatically with increasing A/σ. If $A/\sigma = 2.0$, for instance, $P_e \approx 2 \times 10^{-2}$ while if $A/\sigma = 4.0$, $P_e \approx 3 \times 10^{-5}$. Incidentally, most applications require error probabilities of order 10^{-4} or less.

Because we have taken due care in our examination of errors for binary signals, the extension to multilevel or μ-ary signals is quite straightforward, providing the noise is gaussian. Consider, for instance, a polar trinary signal ($\mu = 3$) with output pulse amplitudes $K_R a_k = +2A$, 0, or $-2A$, representing the trinary digits **2**, **1**, and **0**, respectively. The equivalent to Fig. 4.26 has three gaussian density functions (Fig. 4.27), and we see that two threshold levels are required. Under the usual condition of equiprobable digits, the optimum threshold levels are easily shown to be $\pm A$, for which

$$P_{e_2} = P_{e_0} = Q\left(\frac{A}{\sigma}\right)$$

whereas

$$P_{e_1} = 2Q\left(\frac{A}{\sigma}\right)$$

because both positive and negative noise excursions cause errors when $K_R a_k = 0$. (By another choice of thresholds it is possible to equalize the per-digit error probabilities, but the cost is increased net probability P_e.) Hence

$$P_e = P_2 P_{e_2} + P_1 P_{e_1} + P_0 P_{e_0} = \frac{4}{3} Q\left(\frac{A}{\sigma}\right)$$

where we have inserted $P_2 = P_1 = P_0 = 1/3$.

Generalizing to arbitrary μ with equal digit probabilities, similar reasoning gives

$$P_e = 2\left(1 - \frac{1}{\mu}\right) Q\left(\frac{A}{\sigma}\right) \tag{9}$$

where the *spacing* between adjacent output pulse amplitudes is $2A$—so $2(\mu - 1)A$ is the peak-to-peak range—and the $\mu - 1$ threshold levels are centered between the pulse amplitudes. Equation (9) clearly reduces to Eq. (8*b*) when $\mu = 2$. Nonetheless it fails to tell the full story, for three reasons: first, a μ-ary digit in general represents more information than a binary digit; second, there are differing severities of error in μ-ary systems, depending on whether the noise shifts the apparent amplitude by one or more steps; third, the relationship between error probability and signal-to-noise ratio is not explicit. An analytic assessment of the first two is quite difficult, but the third can be treated as follows.

Signal-to-Noise Ratios

If the contaminating noise is white, then the filtered noise power is

$$N = \eta B \tag{10a}$$

and

$$\sigma = \sqrt{N} = \sqrt{\eta B} \tag{10b}$$

since we have assumed $\bar{n} = 0$. If the noise is not white, the best receiving filter is not an ideal LPF. The question of *optimum* filtering for digital transmission with arbitrary noise spectrum is covered in Chap. 10.

To calculate the signal power, a short digression on the properties of $p(t) = \text{sinc}\, rt$ is needed. It is readily shown (Prob. 4.32) that

$$\int_{-\infty}^{\infty} A_k p\left(t - \frac{k}{r}\right) A_m p\left(t - \frac{m}{r}\right) dt = \begin{cases} \dfrac{A_k^2}{r} & m = k \\ 0 & m \neq k \end{cases} \tag{11}$$

which means that the pulses that make up $K_R x(t)$ are *mutually orthogonal* with energy A_k^2/r per pulse. But orthogonality is a sufficient condition for *superposition of energy*, and a message M digits long, say

$$K_R x(t) = \sum_{k=1}^{M} A_k p\left(t - \frac{k}{r}\right)$$

has energy

$$E_M = \frac{1}{r} \sum_{k=1}^{M} A_k^2$$

which would also be true for *rectangular* pulses with $\tau = 1/r$. Then, since average power equals energy per unit time and an M-digit message is M/r seconds long, the received signal power is

$$S_R = \frac{E_M}{(M/r)} = \frac{1}{M} \sum_{k=1}^{M} A_k^2$$

or simply the *average* of A_k^2. Generalizing, we have that if

$$K_R x(t) = \sum_{k=-\infty}^{\infty} A_k \operatorname{sinc}(rt - k) \qquad (12a)$$

and if the A_k are statistically independent, then

$$S_R = K_R^2 \overline{x^2} = \overline{A^2} \qquad (12b)$$

where $\overline{A^2}$ is the statistical or *ensemble average* of A_k^2. As it happens, this result will also be of use several times in future chapters.

In the case of a polar binary wave with $A_k = \pm A$, equally likely,

$$\overline{A^2} = (+A)^2 P(+A) + (-A)^2 P(-A) = \frac{A^2}{2} + \frac{A^2}{2} = A^2$$

In the polar μ-ary case with μ even,

$$A_k = \pm A, \pm 3A, \ldots, \pm(\mu - 1)A$$

and assuming equiprobable symbols so $P(A_k) = 1/\mu$, applying Eq. (1), Sect. 3.3, gives

$$\overline{A^2} = \sum_{k=1}^{\mu} A_k^2 P(A_k) = \frac{2}{\mu} \sum_{k=1}^{\mu/2} (2k-1)^2 A^2$$

Thus, with the help of the summations in Table B, we have

$$\overline{A^2} = \frac{\mu^2 - 1}{3} A^2 \qquad (13)$$

Equation (13) also holds for any polar wave, μ even or odd, with uniform amplitude spacing $2A$.

Finally, combining Eqs. (10), (12), and (13), the signal-to-noise ratio at the input to the A/D converter is

$$\frac{S}{N} = \frac{S_R}{\eta B} = \frac{\mu^2 - 1}{3} \left(\frac{A}{\sigma}\right)^2 \qquad (14)$$

and Eq. (9) can be written as

$$P_e = 2\left(1 - \frac{1}{\mu}\right) Q\left(\sqrt{\frac{3}{\mu^2 - 1} \frac{S}{N}}\right) \qquad (15)$$

Alternatively, it is useful to define the system parameter

$$\rho \triangleq \frac{S_R}{\eta r} \qquad (16)$$

where r is the signaling rate and, hence, S_R/r equals the average received energy per digit. The parameter ρ plays essentially the same role in digital transmission that γ plays in analog transmission. For the pulse shape in question, $r = 2B$ so $S_R/\eta B = 2S_R/\eta r = 2\rho$ and Eq. (15) becomes

$$P_e = 2\left(1 - \frac{1}{\mu}\right) Q\left(\sqrt{\frac{6\rho}{\mu^2 - 1}}\right) \qquad (17a)$$

$$= Q(\sqrt{2\rho}) \qquad \mu = 2 \qquad (17b)$$

Again, like analog transmission, Eq. (15) or (17) is an upper bound on system performance—that is, a *lower bound* on P_e—and various imperfections will cause the error probability to be higher than predicted. In particular, the wasted DC power in a *unipolar* waveform changes the picture appreciably.

EXERCISE 4.6 Consider a unipolar binary system with $A_k = 2A$, or 0, and the decision threshold at A. Show that $S_R = 2A^2$ if $P_1 = P_0$, and

$$P_e = Q\left(\frac{A}{\sigma}\right) = Q\left(\sqrt{\frac{S}{2N}}\right) = Q(\sqrt{\rho}) \qquad (18)$$

Then calculate P_e for a polar and unipolar system, both having $\rho = 8.0$. *Ans.*: 3×10^{-5}, 2×10^{-3}.

Example 4.2

It is desired to transmit quarternary ($\mu = 4$) digits at a rate of 5,000 per second on a system having $S_R/\eta = 80{,}000$. Assuming sinc pulses are used, the bandwidth required is $B = r/2 = 2{,}500$, and applying Eqs. (16) and (17) gives

$$\rho = \frac{S_R}{\eta r} = 16.0$$

$$P_e = \frac{3}{2} Q\left(\sqrt{\frac{2\rho}{5}}\right) \approx 10^{-2}$$

so the system is not very reliable.

Suppose, however, that each quaternary digit is replaced by *two* binary digits. The required rate is then $r' = 2 \times 5{,}000 = 10{,}000$ and

$$\rho' = \frac{S_R}{\eta r'} = 8.0$$

$$P_e' = Q(\sqrt{2\rho'}) \approx 3 \times 10^{-5}$$

which is a substantial performance improvement. The price of that improvement is a *larger bandwidth*, $B' = 2B$, plus *encoding* and *decoding* units.

Later chapters will generalize these observations under the headings of *wideband noise reduction* and *information theory*. ////

Regenerative Repeaters

Long-haul transmission requires repeaters, be it for analog or digital communication. But unlike analog-message repeaters, digital repeaters can be *regenerative* in the sense that each repeater has an A/D converter as well as an amplifier. If the error probability per repeater is reasonably low and the number of hops M is large, the regeneration advantage turns out to be rather spectacular. This will be demonstrated for the case of polar binary transmission.

When analog repeaters are used and Eq. (17), Sect. 4.1, applies, the final signal-to-noise ratio is $S/N = (1/M)(S/N)_1$ and

$$P_e = Q\left[\sqrt{\frac{1}{M}\left(\frac{S}{N}\right)_1}\right] \qquad (19)$$

where $(S/N)_1$ is the signal-to-noise ratio after one hop. Therefore, the transmitted power *per repeater* must be increased linearly with M just to stay even, a factor not to be sneezed at since, for example, it takes 100 or more repeaters to cross the continent. The $1/M$ term in Eq. (19) stems from the fact that the contaminating noise progressively builds up from repeater to repeater.

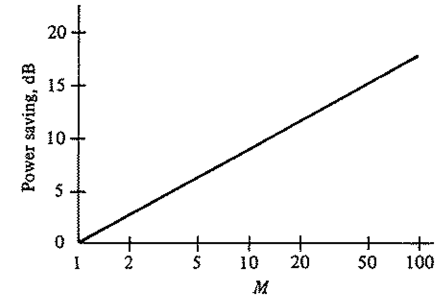


FIGURE 4.28
Power saving gained by regeneration as a function of the number of repeaters, $P_e = 10^{-5}$

In contrast, a regenerative repeater station consists of a complete receiver and transmitter back to back in one package. The receiving portion converts incoming signals to message digits, making a few errors in the process; the digits are then delivered to the transmitting portion, which in turn generates a new signal for transmission to the next station. The regenerated signal is thereby completely stripped of random noise but does contain some errors.

To analyze the performance, let ϵ be the error probability at each repeater, namely,

$$\epsilon = Q\left[\sqrt{\left(\frac{S}{N}\right)_1}\right] \qquad (20)$$

assuming identical units. As a given digit passes from station to station, it may suffer cumulative conversion errors. If the number of erroneous conversions is *even*, they cancel out, and a correct digit is delivered to the destination. (Note that this is true only for *binary* data.) The probability of n errors in M successive conversions is given by the *binomial distribution* of Eq. (1), Sect. 3.4:

$$P_M(n) = \binom{M}{n} \epsilon^n (1-\epsilon)^{M-n}$$

The net error probability is then the probability that n is *odd*; specifically,

$$P_e = \sum_{n\ \mathrm{odd}} P_M(n) = \binom{M}{1}\epsilon(1-\epsilon)^{M-1} + \binom{M}{3}\epsilon^3(1-\epsilon)^{M-3} + \cdots \approx M\epsilon$$

where the approximation applies for $\epsilon \ll 1$ and M not too large. Hence, inserting Eq. (20),

$$P_e \approx MQ\left[\sqrt{\left(\frac{S}{N}\right)_1}\right] \qquad (21)$$

so P_e increases linearly with M, which generally requires a much smaller power increase to counteract than Eq. (19).

Figure 4.28 illustrates the power saving provided by regeneration as a function of M, the error probability being fixed at $P_e = 10^{-5}$. Thus, for example, a 10-station nonregenerative baseband system requires about 8.5 dB more transmitted power (per repeater) than a regenerative system.

4.6 PROBLEMS

4.1 (Sect. 4.1) A 20-km cable system has $S_T = 10$ dBW, $\alpha = 0.64$, and $B_N = 500$ kHz. The noise is additive and white with $\mathcal{T}_N = 25\mathcal{T}_0$. Find $(S/N)_D$ and $\mathcal{G}_R$ in decibels such that $S_D = 0$ dBW. *Ans.*: 33, 100.

4.2 (Sect. 4.1) Repeat Prob. 4.1 with $\alpha = 0.32$ and $B_N = 5$ MHz.

4.3 (Sect. 4.1) A 400-km repeater system employs a cable with 8.68 $\alpha = 0.5$. If $S_T/\eta_1 B_N = 70$ dB, what is the minimum number of repeaters such that $(S/N)_D \geq 20$ dB? Assume equal repeater spacing so $\mathcal{L}_1 = \mathcal{L}^{1/M}$.

4.4★ (Sect. 4.1) All other parameters being fixed, show that the number of equally spaced repeaters that maximizes $(S/N)_D$ is $M = \ln \mathcal{L} = 0.23\mathcal{L}_{dB}$.

4.5★ (Sect. 4.1) Owing to nonlinearities, the output of a baseband system is $y_D(t) = x(t) + n(t) + \frac{1}{2}[x(t) + n(t)]^2$. Find $\overline{y_D^2}$ assuming that $x(t)$ and $n(t)$ are independent and all of their odd moments are zero. Is it possible to define $(S/N)_D$ in this case?

4.6★ (Sect. 4.1) The open-circuit signal output of an oscillator is $A \cos 2\pi f_0 t$. The source resistance is R_S and there is internally generated thermal noise with temperature $\mathcal{T}_N$. A capacitor C is placed across the output terminals to improve S/N.

(*a*) Obtain an expression for S/N.

(*b*) What value of C maximizes S/N?

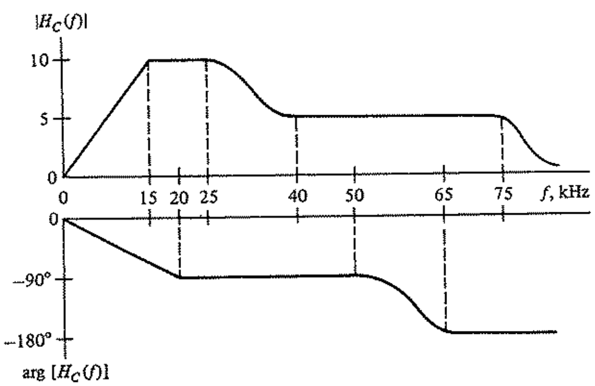


FIGURE P4.1.

4.7 (Sect. 4.2) Figure P4.1 shows the amplitude ratio and phase shift of a certain transmission channel. What frequency range or ranges has: amplitude distortion, phase distortion, distortionless transmission?

4.8 (Sect. 4.2) Show that an RC LPF gives essentially distortionless transmission if $x(t)$ is bandlimited in $W \ll B = 1/2\pi RC$.

4.9 (Sect. 4.2) The input to an RC LPF is the test signal in Fig. 4.4. Plot the output waveform when $f_0 = B/3$.

4.10 (Sect. 4.2) Find $t_d(f)$ for an RC LPF and evaluate it at $f = B/4$, B, and $4B$ when $B = 1$ kHz.

4.11 (Sect. 4.2) Consider a transfer function with *ripples in the amplitude ratio*, $H(f) = (1 + 2\alpha \cos \omega t_0)e^{-j\omega t_d}$, $|\alpha| \leq \frac{1}{2}$.

(*a*) Show that $y(t) = x(t - t_d) + \alpha x(t - t_d + t_0) + \alpha x(t - t_d - t_0)$, so there is a pair of "echoes".

(*b*) Taking $\alpha = \frac{1}{2}$ and $x(t) = \Pi(t/\tau)$, sketch $y(t)$ for $t_0 = 2\tau$, $\tau/2$, and $\tau/4$.

4.12★ (Sect. 4.2) Find $y(t)$ in terms of $x(t)$ when there are *small ripples in the phase shift*, i.e., $H(f) = \exp[-j(\omega t_d - \alpha \sin \omega t_0)]$, $|\alpha| \ll \pi$. Compare with Prob. 4.11. (*Hint*: Use a series expansion for $\exp(j\alpha \sin \omega t_0)$.)

4.13 (Sect. 4.2) Sketch $|H_{eq}(f)|$ and $\arg[H_{eq}(f)]$ needed to equalize $H_C(f)$ in Fig. P4.1 over $5 < |f| < 25$ kHz.

4.14 (Sect. 4.2) Suppose $x(t) = v(t) + \cos 2\pi f_1 t$ is applied to a nonlinear system with $y(t) = x(t) + 0.4x^2(t) + 0.1x^3(t)$.

(*a*) Find $y(t)$ and sketch a typical $Y(f)$ when $v(t)$ is bandlimited in $W \ll f_1$.

(*b*) If $v(t) = \cos 2\pi f_2 t$, $f_2 > f_1$, list all the frequency components in $y(t)$.

4.15 (Sect. 4.3) A system designed for telephone-quality voice transmission has $(S/N)_D = 30$ dB when $S_T = -3$ dBW. If the bandwidth is appropriately increased, what value of S_T is required to upgrade the system for high-fidelity audio transmission, all other factors being unchanged? *Ans.*: 30 to 40 dBW.

4.16 (Sect. 4.3) A system designed for an analog signal with $W = 10$ kHz uses an RC LPF with 3-dB bandwidth $B = 15$ kHz at the receiver.

(*a*) Find $(S/N)_D$ in terms of γ.

(*b*) Repeat for a second-order Butterworth filter (Exercise 3.10, Sect. 3.6) with $B = 12$ kHz.

4.17 (Sect. 4.3) A distorting channel with $H_C(f) = [1 + j(2f/W)]^{-1}$ and white noise is used for a signal with $G_x(f) = (1/2W)\Pi(f/2W)$. To compensate for the distortion, the receiving filter is $H_R(f) = [1 + j(2f/W)]\Pi(f/2W)$. The transmitting filter is simply an amplifier with unit gain. Show that $(S/N)_D \approx 0.78\gamma$.

4.18★ (Sect. 4.3) Referring to Fig. 4.15, obtain expressions for $\mathcal{G}_T$, $\mathcal{L}$, and $\mathcal{G}_R$ in terms of $H_T(f)$, $H_C(f)$, $H_R(f)$, and $G_x(f)$. (*Hint*: By definition, $\mathcal{G}_T = S_T/\overline{x^2}$, etc.)

4.19★ (Sect. 4.3) Given $H_C(f)$ and $G_x(f)$ in Prob. 4.17, find the optimum terminal filters and evaluate $(S/N)_{D_{max}}$ in terms of γ.

4.20 (Sect. 4.4) Redraw Fig. 4.18 for the case of an RC LPF with $B = 1/2\tau$ and $B = 1/4\tau$.

4.21 (Sect. 4.4) A sinc pulse $x(t) = A \operatorname{sinc} 2Wt$ is applied to an ideal LPF having bandwidth B. Taking the duration of sinc at to be $\tau \triangleq 2/a$, plot the ratio of output duration to input duration as a function of B/W.

4.22 (Sect. 4.4) A rectangular pulse with $\tau \gg 1/B$ is applied to an RC LPF. Find the 10 to 90 percent rise time in terms of the 3-dB bandwidth B.