CHAPTER

ELEVEN

BASEBAND DIGITAL TRANSMISSION

This chapter launches our study of digital communication systems. We focus initially on baseband transmission in order to emphasize the generic concepts and problems of digital communication, with or without carrier modulation, and to bring out the differences between digital and analog transmission. Following an overview of digital signals and systems, we'll analyze the limitations imposed by additive noise and transmission bandwidth. We'll also look at practical design considerations such as regenerative repeaters, equalization, and synchronization.

11.1 DIGITAL SIGNALS AND SYSTEMS

Fundamentally, a digital message is nothing more than an *ordered sequence of symbols* produced by a discrete information source. The source draws from an *alphabet* of $M \geq 2$ different symbols, and produces output symbols at some average *rate* r. For instance, a typical computer terminal has an alphabet of $M \approx 90$ symbols, equal to the number of character keys multiplied by two to account for the shift key. When you operate the terminal as fast as you can, you become a discrete information source producing a digital message at a rate of perhaps $r \approx 5$ symbols per second. The computer itself works with just $M = 2$ internal symbols, represented by LOW and HIGH electrical states. We usually associate these two symbols with the *binary digits* 0 and 1, known as *bits* for short. Data transfer rates within a computer may exceed $r = 10^8$.

The task of a digital communication system is to transfer a digital message from the source to the destination. But finite transmission bandwidth sets an upper limit to the symbol rate, and noise causes errors to appear in the output message. Thus, *signaling rate* and *error probability* play roles in digital communication similar to those of bandwidth and signal-to-noise ratio in analog communication. As preparation for the analysis of signaling rate and error probability, we must first develop the description and properties of digital signals.

Digital PAM Signals

Digital message representation at baseband commonly takes the form of an *amplitude-modulated pulse train.* We express such signals by writing

$$x(t) = \sum_k a_k p(t - kD) \tag{1}$$

where the modulating amplitude a_k represents the kth symbol in the message sequence, so the amplitudes belong to a set of M discrete values. The index k ranges from $-\infty$ to $+\infty$ unless otherwise stated. Equation (1) defines a digital PAM signal, as distinguished from those rare cases when pulse-duration or pulse-position modulation is used for digital transmission.

The unmodulated pulse $p(t)$ may be rectangular or some other shape, subject to the conditions

$$p(t) = \begin{cases} 1 & t = 0 \\ 0 & t = \pm D, \pm 2D, \ldots \end{cases} \tag{2}$$

This condition ensures that we can recover the message by sampling $x(t)$ periodically at $t = KD$, $K = 0, \pm 1, \pm 2, \ldots$, since

$$x(KD) = \sum_k a_k p(KD - kD) = a_K$$

The rectangular pulse $p(t) = \Pi(t/\tau)$ satisfies Eq. (2) if $\tau \leq D$, as does any time-limited pulse with $p(t) = 0$ for $|t| \geq D/2$.

Note that D does not necessarily equal the pulse duration but rather the pulse-to-pulse interval or the time allotted to one symbol. Thus, the *signaling rate* is

$$r \triangleq 1/D \tag{3a}$$

measured in symbols per second or *baud.* In the special but important case of *binary* signaling ($M = 2$), we write $D = T_b$ for the bit duration and the bit rate is†

$$r_b = 1/T_b \tag{3b}$$

measured in bits per second, abbreviated bps or b/s. The notation T_b and r_b will be used to identify results that apply only for binary signaling.

† The more common notation R for bit rate risks confusion with autocorrelation functions and with information rate defined in Chap. 15.

Figure 11.1-1 depicts various PAM *formats* for the binary message 10110100, taking rectangular pulses for clarity. The simple *on-off* waveform in part *a* represents each 0 by an "off" pulse ($a_k = 0$) and each 1 by an "on" pulse with amplitude $a_k = A$ and duration $T_b/2$ followed by a return to the zero level. We therefore call this a *return-to-zero* (RZ) format. A *nonreturn-to-zero* (NRZ) format has "on" pulses for full bit duration T_b, as indicated by the dashed lines. Internal computer waveforms are usually of this type. The NRZ format puts more energy into each pulse, but requires synchronization at the receiver because there's no separation between adjacent pulses.

The unipolar nature of an on-off signal results in a *DC component* that carries no information and wastes power. The *polar* signal in part *b* has opposite polarity pulses, either RZ or NRZ, so its DC component will be zero if the message contains 1s and 0s in equal proportion. This property also applies to the *bipolar*

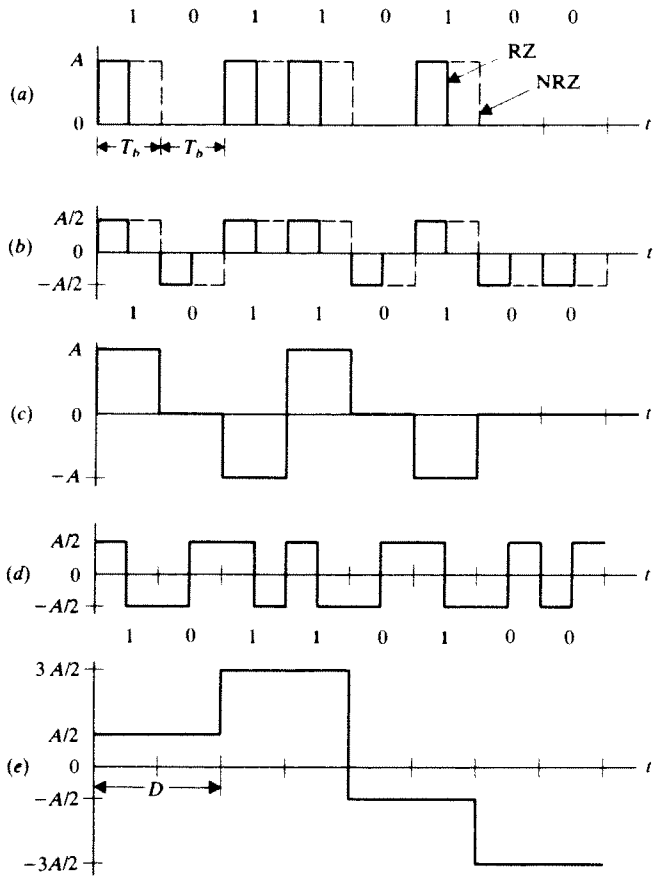


Figure 11.1-1 Binary PAM formats with rectangular pulses. (*a*) Unipolar RZ and NRZ; (*b*) polar RZ and NRZ; (*c*) bipolar NRZ; (*d*) split-phase Manchester; (*e*) polar quaternary NRZ.

Table 11.1-1

a_k	Natural code	Gray code
$3A/2$	11	10
$A/2$	10	11
$-A/2$	01	01
$-3A/2$	00	00

signal in part *c*, where successive 1s are represented by pulses with alternating polarity. The bipolar format, also known as *pseudo-trinary* or *alternate mark inversion* (AMI), eliminates ambiguities that might be caused by transmission *sign inversions*—a problem characteristic of switched telephone links.

The *split-phase Manchester* format in part *d* represents 1s with a positive half-interval pulse followed by a negative half-interval pulse, and vice versa for the representation of 0s. This format is also called *twinned binary*. It guarantees zero DC component regardless of the message sequence. However, it requires an absolute sense of polarity at the receiver.

Finally, Fig. 11.1-1*e* shows a *quaternary* signal derived by grouping the message bits in blocks of two and using four amplitude levels to prepresent the four possible combinations 00, 01, 10, and 11. Thus, $D = 2T_b$ and $r = r_b/2$. Different assignment rules or *codes* may relate the a_k to the grouped message bits. Two such codes are listed in Table 11.1-1. The Gray code has advantages relative to noise-induced errors because only one bit changes going from level to level.

Quaternary coding generalizes to *M-ary* coding in which blocks of n message bits are represented by an M-level waveform with

$$M = 2^n \tag{4a}$$

Since each pulse now corresponds to $n = \log_2 M$ bits, the M-ary signaling rate has been decreased to

$$r = \frac{r_b}{\log_2 M} \tag{4b}$$

But increased signal power would be required to maintain the same spacing between amplitude levels.

Transmission Limitations

Now consider the linear baseband transmission system diagrammed in Fig. 11.1-2*a*. We'll assume for convenience that the transmitting amplifier compensates for the transmission loss, and we'll lump any interference together with the additive noise. After lowpass filtering to remove out-of-band contaminations, we have the signal-plus-noise waveform

$$y(t) = \sum_k a_k \tilde{p}(t - t_d - kD) + n(t)$$

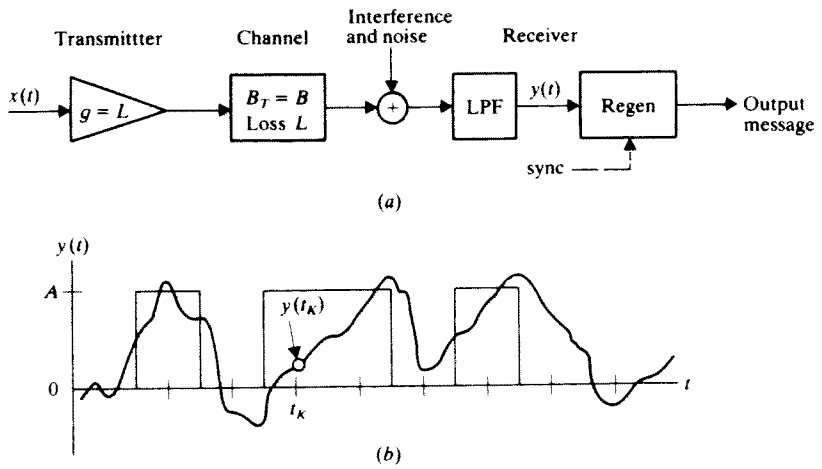


Figure 11.1-2 (*a*) Baseband transmission system; (*b*) signal-plus-noise waveform.

where t_d is the transmission delay and $\tilde{p}(t)$ stands for the pulse shape with transmission distortion. Figure 11.1-2*b* illustrates what $y(t)$ might look like when $x(t)$ is the unipolar NRZ signal in Fig. 11.1-1*a*.

Recovering the digital message from $y(t)$ is the task of the *regenerator*. An auxiliary synchronization signal may help the regeneration process by identifying the optimum sampling times

$$t_K = KD + t_d$$

If $\tilde{p}(0) = 1$ then

$$y(t_K) = a_K + \sum_{k \neq K} a_k \tilde{p}(KD - kD) + n(t_K) \tag{5}$$

whose first term is the desired message information. The last term of Eq. (5) is the noise contamination at t_K, while the middle term represents cross talk or spillover from other signal pulses—a phenomenon given the descriptive name *intersymbol interference* (ISI). The combined effects of noise and ISI may result in *errors* in the regenerated message. For instance, at the sample time t_K indicated in Fig. 11.1-2*b*, $y(t_K)$ is closer to 0 even though $a_K = A$.

We know that if $n(t)$ comes from a white-noise source, then its mean square value can be reduced by reducing the bandwidth of the LPF. We also know that lowpass filtering causes pulses to spread out, which would increase the ISI. Consequently, a fundamental limitation of digital transmission is the relationship between ISI, bandwidth, and signaling rate.

This problem emerged in the early days of telegraphy, and Harry Nyquist (1924, 1928) first stated the relationship as follows:

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

The condition $r \leq 2B$ agrees with our pulse-resolution rule $B \geq 1/2\tau_{min}$ in Sect. 3.4 if we require $p(t)$ to have duration $\tau \leq D = 1/r$.

The second part of Nyquist's statement is easily proved by assuming that $r = 2(B + \epsilon) > 2B$. Now suppose that the message sequence happens to consist of two symbols alternating forever, such as 101010... The resulting waveform $x(t)$ then is *periodic* with period $2D = 2/r$ and contains only the fundamental frequency $f_0 = B + \epsilon$ and its harmonics. Since no frequency greater than B gets through the channel, the output signal will be zero—aside from a possible but useless DC component.

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

$$p(t) = \text{sinc } rt = \text{sinc } t/D \tag{6a}$$

having the *bandlimited* spectrum

$$P(f) = \mathcal{F}[p(t)] = \frac{1}{r}\,\Pi\left(\frac{f}{r}\right) \tag{6b}$$

Since $P(f) = 0$ for $|f| > r/2$, this pulse suffers no distortion from an ideal lowpass frequency response with $B \geq r/2$ and we can take $r = 2B$. Although $p(t)$ is not timelimited, it does have *periodic zero crossings* at $t = \pm D, \pm 2D, \ldots$, which satisfies Eq. (2). (See Fig. 10.1-6 for an illustration of this property.) Nyquist also derived other bandlimited pulses with periodic zero crossings spaced by $D > 1/2B$ so $r < 2B$, a topic we set aside to pick up again in Sect. 11.3 after discussing noise and errors in Sect. 11.2.

Meanwhile, note that any real channel needs *equalization* to approach an ideal frequency response. Such equalizers often require experimental adjustment in the field because we don't know the channel characteristics exactly. An important experimental display is the so-called *eye pattern*, which further clarifies digital transmission limitations.

Consider the distorted but noise-free polar binary signal in Fig. 11.1-3*a*. When displayed on a long-persistence oscilloscope with appropriate synchronization and sweep time, we get the superposition of successive symbol intervals shown in Fig. 11.1-3*b*. The shape of this display accounts for the name "eye pattern." A distorted M-ary signal would result in $M - 1$ "eyes" stacked vertically.

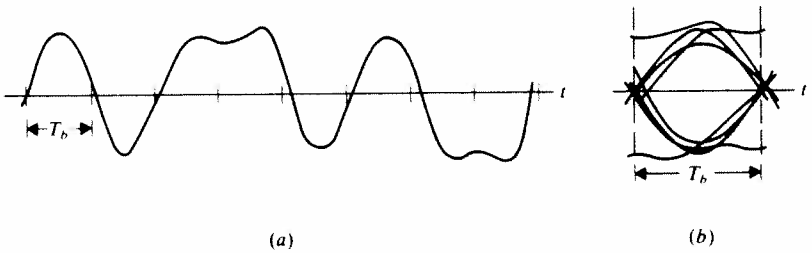


Figure 11.1-3 (*a*) Distorted polar binary signal; (*b*) eye pattern.

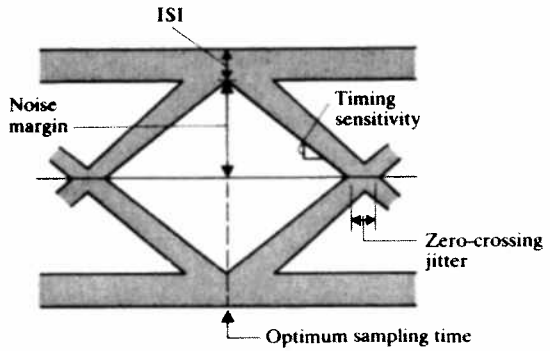


Figure 11.1-4 Generalized binary eye pattern.

Figure 11.1-4 represents a generalized binary eye pattern with labels identifying significant features. The optimum sampling time corresponds to the maximum eye opening. ISI at this time partially closes the eye and thereby reduces the *noise margin*. If synchronization is derived from the zero crossings, as it usually is, zero-crossing distortion produces *jitter* and results in nonoptimum sampling times. The slope of the eye pattern in the vicinity of the zero crossings indicates the sensitivity to timing error. Finally, any nonlinear transmission distortion would reveal itself in an asymmetric or "squinted" eye.

Exercise 11.1-1 Determine the relation between r and B when $p(t) = \text{sinc}^2 at$.

Power Spectra of Digital PAM

The pulse spectrum $P(f) = \mathcal{F}[p(t)]$ provides some hints about the *power spectrum* of a digital PAM signal $x(t)$. If $p(t) = \text{sinc } rt$, as a case in point, then $P(f)$ in Eq. (6*b*) implies that $G_x(f) = 0$ for $|f| > r/2$. However, detailed knowledge of $G_x(f)$ provides additional and valuable information relative to digital transmission.

A simplified random digital wave with $p(t) = \Pi(t/D)$ was tackled in Chap. 5. Under the conditions

$$E[a_k a_i] = \begin{cases} \sigma_a^2 & i = k \\ 0 & i \neq k \end{cases}$$

we found that $G_x(f) = \sigma_a^2 D \text{ sinc}^2 fD$. Now, substituting $P(f) = D \text{ sinc } fD$, we write

$$G_x(f) = \frac{\sigma_a^2}{D}\,|P(f)|^2 \tag{7}$$

This expression holds for *any* digital PAM signal with pulse spectrum $P(f)$ when the a_k are uncorrelated and have zero mean value.

But unipolar signal formats have $\overline{a_k} \neq 0$ and, in general, we can't be sure that the message source produces uncorrelated symbols. A more realistic approach

therefore models the source as a *discrete stationary random process.* Ensemble averages of the a_k are then given by the autocorrelation function

$$R_a(n) = E[a_k a_{k-n}] \tag{8}$$

analogous to writing $R_v(\tau) = E[v(t)v(t-\tau)]$ for a stationary random signal $v(t)$. The integers n and k in Eq. (8) reflect the time-discrete nature of a digital sequence.

If a digital PAM signal $x(t)$ has the pulse spectrum $P(f)$ and amplitude autocorrelation $R_a(n)$, its power spectrum is

$$G_x(f) = \frac{1}{D}|P(f)|^2 \sum_{n=-\infty}^{\infty} R_a(n)e^{-j2\pi nfD} \tag{9}$$

Despite the formidable appearance of Eq. (9), it easily reduces to Eq. (7) when $R_a(0) = \sigma_a^2$ and $R_a(n) = 0$ for $n \neq 0$. In the case of uncorrelated message symbols but $\overline{a_k} = m_a \neq 0$,

$$R_a(n) = \begin{cases} \sigma_a^2 + m_a^2 & n = 0 \\ m_a^2 & n \neq 0 \end{cases} \tag{10}$$

and

$$\sum_{n=-\infty}^{\infty} R_a(n)e^{-j2\pi nfD} = \sigma_a^2 + m_a^2 \sum_{n=-\infty}^{\infty} e^{-j2\pi nfD}$$

Then, drawing upon Poisson's sum formula,

$$\sum_{n=-\infty}^{\infty} e^{-j2\pi nfD} = \frac{1}{D}\sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{D}\right)$$

and therefore

$$G_x(f) = \sigma_a^2 r|P(f)|^2 + (m_a r)^2 \sum_{n=-\infty}^{\infty} |P(nr)|^2\, \delta(f - nr) \tag{11}$$

Here we have inserted $r = 1/D$ and used the sampling property of impulse multiplication.

The important result in Eq. (11) reveals that the power spectrum of $x(t)$ contains impulses at *harmonics* of the signaling rate r, unless $m_a = 0$ or $P(f) = 0$ at all $f = nr$. Hence, a synchronization signal can be obtained by applying $x(t)$ to a narrow BPF (or PLL filter) centered at one of these harmonics. We can also calculate the total *average power* $\overline{x^2}$ by integrating $G_x(f)$ over all f. Thus,

$$\overline{x^2} = \sigma_a^2 rE_p + (m_a r)^2 \sum_{n=-\infty}^{\infty} |P(nr)|^2 \tag{12}$$

where E_p equals the energy in $p(t)$. For Eq. (12), and hereafter, we presume the conditions in Eq. (10) barring information to the contrary.

The derivation of Eq. (9) starts with the definition of power spectrum from Eq. (9), Sect. 5.2. Specifically, we write

$$G_x(f) \triangleq \lim_{T\to\infty} \frac{1}{T} E[|X_T(f)|^2]$$

in which $X_T(f)$ is the Fourier transform of a truncated sample function $x_T(t) = x(t)$ for $|t| < T/2$. Next, let $T = (2K + 1)D$ so the limit $T \to \infty$ corresponds to $K \to \infty$. Then, for $K \gg 1$,

$$x_T(t) = \sum_{k=-K}^{K} a_k\, p(t - kD)$$

$$X_T(f) = \sum_{k=-K}^{K} a_k\, P(f)e^{-j\omega kD}$$

and

$$|X_T(f)|^2 = |P(f)|^2 \left(\sum_{k=-K}^{K} a_k\, e^{-j\omega kD}\right)\left(\sum_{i=-K}^{K} a_i\, e^{+j\omega iD}\right)$$

After interchanging the order of expectation and summation we have

$$E[|X_T(f)|^2] = |P(f)|^2 \rho_K(f)$$

with

$$\rho_K(f) = \sum_{k=-K}^{K}\sum_{i=-K}^{K} E[a_k a_i]e^{-j\omega(k-i)D}$$

where $E[a_k a_i] = R_a(k - i)$.

The double summation for $\rho_K(f)$ can be manipulated into the single sum

$$\rho_K(f) = (2K + 1)\sum_{n=-2K}^{2K}\left(1 - \frac{|n|}{2K+1}\right)R_a(n)e^{-j\omega nD}$$

Substituting these expressions in the definition of $G_x(f)$ finally gives

$$G_x(f) = \lim_{K\to\infty} \frac{1}{(2K+1)D}|P(f)|^2\rho_K(f)$$

$$= \frac{1}{D}|P(f)|^2 \sum_{n=-\infty}^{\infty} R_a(n)e^{-j\omega nD}$$

as stated in Eq. (9).

Example 11.1-1 Consider the unipolar binary RZ signal in Fig. 11.1-1*a*, where $p(t) = \Pi(2r_b t)$ so

$$P(f) = \frac{1}{2r_b}\operatorname{sinc}\frac{f}{2r_b}$$

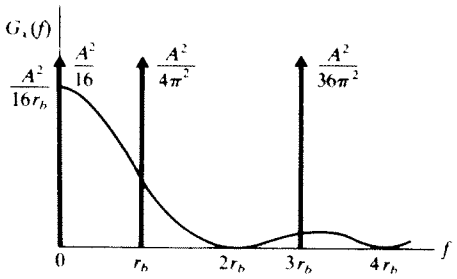


Figure 11.1-5 Power spectrum of unipolar binary RZ signal.

If the source bits are equally likely and statistically independent, then $\overline{a_k} = A/2$, $\overline{a_k^2} = A^2/2$, and Eq. (10) applies with

$$m_a^2 = \sigma_a^2 = \frac{A^2}{4}$$

Using Eq. (11) we find the power spectrum to be

$$G_x(f) = \frac{A^2}{16r_b}\,\text{sinc}^2\,\frac{f}{2r_b} + \frac{A^2}{16}\sum_{n=-\infty}^{\infty}\left(\text{sinc}^2\,\frac{n}{2}\right)\delta(f - nr_b)$$

which is sketched in Fig. 11.1-5 for $f \geq 0$. Notice the absence of impulses at the even harmonics because $P(nr_b) = 0$ when $n = \pm 2, \pm 4, \ldots$.

We could also, in principle, use Eq. (12) to calculate $\overline{x^2}$. However, it should be evident from the waveform that $\overline{x^2} = A^2/4$ when 1s and 0s are equally likely.

Exercise 11.1-2 Modify the results of Example 11.1-1 for $p(t) = \Pi(r_b t)$, corresponding to an NRZ waveform. In particular, show that the only impulse in $G_x(f)$ is at $f = 0$.

Spectral Shaping by Precoding★

Precoding refers to operations that cause the statistics of the amplitude sequence a_k to differ from the statistics of the message sequence. Usually, the purpose of precoding is to *shape* the power spectrum via $R_a(n)$, as distinguished from $P(f)$. To bring out the potential for statistical spectral shaping, we rewrite Eq. (9) in the form

$$G_x(f) = r\,|P(f)|^2\left[R_a(0) + 2\sum_{n=1}^{\infty} R_a(n)\cos(2\pi nf/r)\right] \tag{13}$$

having drawn upon the property $R_a(-n) = R_a(n)$.

Now suppose that $x(t)$ is to be transmitted over a channel having poor low-frequency response—a voice telephone channel perhaps. With appropriate precoding, we can force the bracketed term in Eq. (13) to equal zero at $f = 0$ and

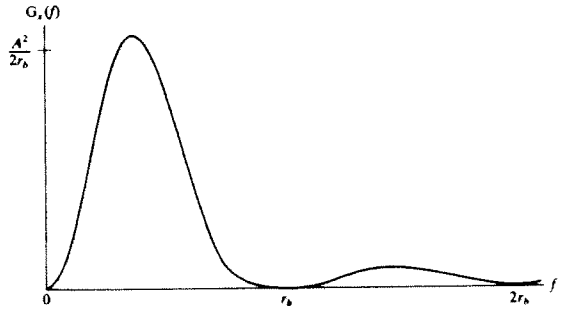


Figure 11.1-6 Power spectrum of bipolar signal.

thereby eliminate any DC component in $G_x(f)$, irrespective of the pulse spectrum $P(f)$. The *bipolar* signal format back in Fig. 11.1-1*c* is, in fact, a precoding technique that removes DC content.

The bipolar signal has three amplitude values, $a_k = +A$, 0, and $-A$. If 1s and 0s are equally likely in the message, then the amplitude probabilities are $P(a_k = 0) = 1/2$ and $P(a_k = +A) = P(a_k = -A) = 1/4$, so the amplitude statistics differ from the message statistics. Furthermore, the assumption of uncorrelated message bits leads to the amplitude correlation

$$R_a(n) = \begin{cases} A^2/2 & n = 0 \\ -A^2/4 & n = 1 \\ 0 & n \geq 2 \end{cases} \tag{14a}$$

Therefore,

$$G_x(f) = r_b\,|P(f)|^2\,\frac{A^2}{2}(1 - \cos 2\pi f/r_b)$$

$$= r_b\,|P(f)|^2 A^2 \sin^2 \pi f/r_b \tag{14b}$$

which is sketched in Fig. 11.1-6 taking $p(t) = \Pi(r_b t)$.

Two other precoding techniques that remove DC content are the split-phase Manchester format (Fig. 11.1-1*d*) and the family of *high density bipolar codes* denoted as HDBn. The HDBn scheme is a bipolar code that also eliminates long signal "gaps" by substituting a special pulse sequence whenever the source produces n successive 0s.

Exercise 11.1-3 Sketch $G_x(f)$ for a bipolar signal with $p(t) = \text{sinc}\, r_b t$. Then use your sketch to show that $\overline{x^2} = A^2/2$.

11.2 NOISE AND ERRORS

Here we investigate noise, errors, and error probabilities in baseband digital transmission, starting with the binary case and generalizing to M-ary signaling.

We assume throughout a distortionless channel, so the received signal is free of ISI. We also assume additive white noise with zero mean value, independent of the signal. (Some of these restrictions will be lifted in the next section.)

Binary Error Probabilities

Figure 11.2-1 portrays the operations of a baseband binary receiver. The received signal plus noise is applied to a lowpass filter whose transfer function has been designed to remove excess noise without introducing ISI. A sample-and-hold (S/H) device triggered at the optimum times extracts from $y(t)$ the sample values

$$y(t_k) = a_k + n(t_k)$$

Comparing successive values of $y(t_k)$ with a fixed *threshold level* V completes the regeneration process. If $y(t_k) > V$, the comparator goes HIGH to indicate a 1; if $y(t_k) < V$, the comparator goes LOW to indicate 0. The regenerator thereby acts as an *analog-to-digital converter*, converting the noisy analog waveform $y(t)$ into a noiseless digital signal $x_e(t)$ with occasional *errors*.

We begin our analysis taking $x(t)$ to be a *unipolar* signal in which $a_k = A$ represents the message bit 1 and $a_k = 0$ represents the message bit 0. Intuitively, then, the threshold should be set at some intermediate level, $0 < V < A$. The regeneration process is illustrated by the waveforms in Fig. 11.2-2. Errors occur when $a_k = 0$ but a positive noise excursion results in $y(t_k) > V$, or when $a_k = A$ but a negative noise excursion results in $y(t_k) < V$.

To formulate the error probabilities, let the *random variable* Y represent $y(t_k)$ at an arbitrary sampling time and let n represent $n(t_k)$. The probability density function of Y obviously involves the noise PDF, but it also depends upon the presence or absence of the signal pulse. We therefore need to work with *conditional* probabilities. In particular, if H_0 denotes the *hypothesis* or assumption that $a_k = 0$ and $Y = n$, we can write the conditional PDF

$$p_Y(y \mid H_0) = p_N(y) \tag{1a}$$

where $p_N(n)$ is the PDF of the noise alone. Similarly, if H_1 denotes the hypothesis that $a_k = A$ and $Y = A + n$, then

$$p_Y(y \mid H_1) = p_N(y - A) \tag{1b}$$

obtained from the linear transformation of $p_N(n)$ with $n = y - A$.

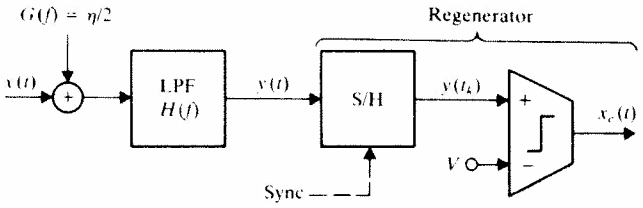


Figure 11.2-1 Baseband binary receiver.

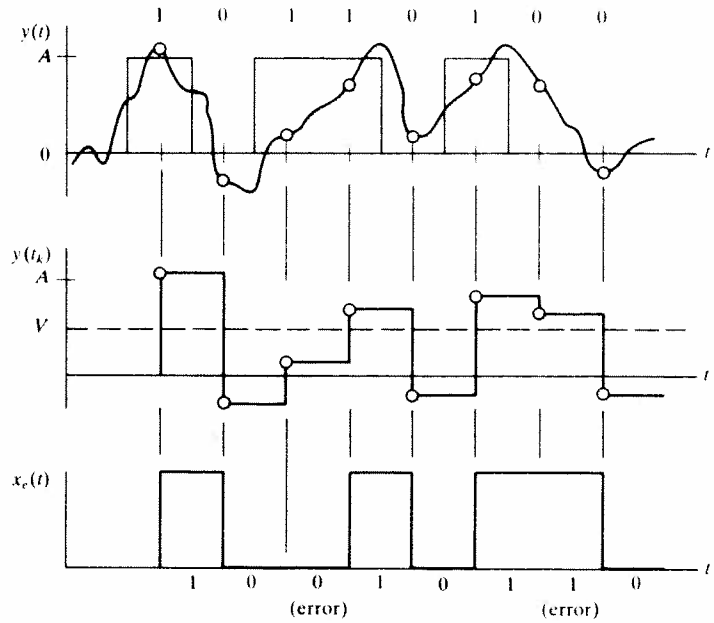


Figure 11.2-2 Regeneration of a unipolar signal. (*a*) Signal plus noise; (*b*) S/H output; (*c*) comparator output.

Figure 11.2-3 shows typical curves of $p_Y(y \mid H_0)$ and $p_Y(y \mid H_1)$ along with a proposed threshold V. The comparator implements the following *decision rule*:

Choose hypothesis H_0 ($a_k = 0$) if $Y < V$

Choose hypothesis H_1 ($a_k = \mathrm{A}$) if $Y > V$

(We ignore the borderline case $Y = V$ whose probability of occurrence will be vanishingly small.) The corresponding regeneration error probabilities are then given by

$$P_{e_0} \triangleq P(Y > V \mid H_0) = \int_V^\infty p_Y(y \mid H_0)\, dy \tag{2a}$$

$$P_{e_1} \triangleq P(Y < V \mid H_1) = \int_{-\infty}^V p_Y(y \mid H_1)\, dy \tag{2b}$$

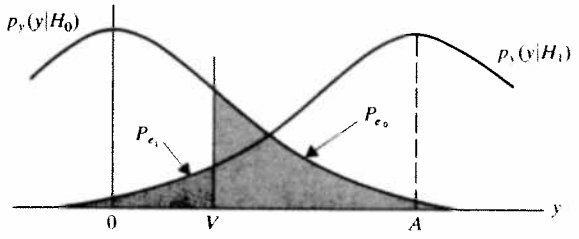


Figure 11.2-3 Conditional PDFs with decision threshold and error probabilities.

equivalent to the shaded areas indicated in Fig. 11.2-3. The area interpretation helps bring out the significance of the threshold level when all other factors remain fixed. Clearly, lowering the threshold reduces P_{e_1} and simultaneously increases P_{e_0}. Raising the threshold has the opposite effect.

But an error in digital transmission is an error, regardless of type. Hence, the threshold level should be adjusted to minimize the *average error probability*

$$P_e = P_0 P_{e_0} + P_1 P_{e_1} \tag{3a}$$

where

$$P_0 = P(H_0) \qquad P_1 = P(H_1) \tag{3b}$$

which stand for the *source digit probabilities*. The *optimum* threshold level V_{opt} must therefore satisfy $dP_e/dV = 0$, and Leibniz's rule for differentiating the integrals in Eq. (2) leads to the general relation

$$P_0\, p_Y(V_{\text{opt}} \mid H_0) = P_1 p_Y(V_{\text{opt}} \mid H_1) \tag{4}$$

But we normally expect 1s and 0s to be *equally likely* in a long string of message bits, so

$$P_0 = P_1 = \tfrac{1}{2} \qquad P_e = \tfrac{1}{2}(P_{e_0} + P_{e_1}) \tag{5a}$$

and

$$p_Y(V_{\text{opt}} \mid H_0) = p_Y(V_{\text{opt}} \mid H_1) \tag{5b}$$

We'll work hereafter with the equally likely condition, unless stated otherwise.

Closer examination of Eq. (5*b*) reveals that V_{opt} corresponds to the point where the conditional PDF curves *intersect*. Direct confirmation of this conclusion is provided by the graphical construction in Fig. 11.2-4 labeled with four relevant areas, α_1 through α_4. The optimum threshold yields $P_{e_1} = \alpha_1 + \alpha_2$, $P_{e_0} = \alpha_3$, and $P_{e_{\min}} = \frac{1}{2}(\alpha_1 + \alpha_2 + \alpha_3)$. A nonoptimum threshold such as $V < V_{\text{opt}}$ yields $P_{e_1} = \alpha_1$ and $P_{e_0} = \alpha_2 + \alpha_3 + \alpha_4$; thus, $P_e = \frac{1}{2}(\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4) = P_{e_{\min}} + \frac{1}{2}\alpha_4 > P_{e_{\min}}$.

Next we make the usual assumption that the noise is *gaussian* with zero mean and variance σ^2, so

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

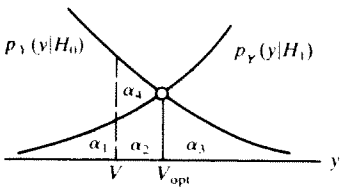


Figure 11.2-4

Substituting this gaussian function into Eqs. (1) and (2) gives

$$P_{e_0} = \int_V^\infty p_N(y)\, dy = Q\left(\frac{V}{\sigma}\right) \tag{6a}$$

$$P_{e_1} = \int_{-\infty}^V p_N(y - A)\, dy = Q\left(\frac{A - V}{\sigma}\right) \tag{6b}$$

where Q is the area under the gaussian tail as previously defined in Fig. 4.4-2. Since $p_N(n)$ has even symmetry, the conditional PDFs $p_Y(y \mid H_0)$ and $p_Y(y \mid H_1)$ intersect at the *midpoint* and $V_{\text{opt}} = A/2$ when $P_0 = P_1 = 1/2$. Furthermore, with $V = V_{\text{opt}}$ in Eq. (6), $P_{e_0} = P_{e_1} = Q(A/2\sigma)$ so the optimum threshold yields *equal digit error probabilities* as well as minimizing the net error probability. Thus, $P_e = \frac{1}{2}(P_{e_0} + P_{e_1}) = P_{e_0} = P_{e_1}$ and

$$P_e = Q\left(\frac{A}{2\sigma}\right) \tag{7}$$

which is the minimum net error probability for binary signaling in gaussian noise when the source digits are equally likely.

Based on Eq. (7), the plot of the Q function in Table T.6 can be interpreted now as a plot of P_e versus $A/2\sigma$. This plot reveals that P_e drops off dramatically when $A/2\sigma$ increases. For instance, $P_e \approx 2 \times 10^{-2}$ at $A/2\sigma = 2.0$ whereas $P_e \approx 10^{-9}$ at $A/2\sigma = 6.0$.

Although we derived Eq. (7) for the case of a unipolar signal, it also holds in the *polar* case if $a_k = \pm A/2$ so the level spacing remains unchanged. The only difference at the receiver is that $V_{\text{opt}} = 0$, midway between the two levels. However, the transmitter needs less signal power to achieve a specified level spacing in the polar signal.

Let's bring out that advantage of polar signaling by expressing A in terms of the average received signal power S_R. If we assume equal digit probabilities and more-or-less rectangular pulses with full-interval duration T_b, then $S_R = A^2/2$ for unipolar signaling while $S_R = A^2/4$ for polar signaling. (These relations should be obvious from the NRZ waveforms in Fig. 11.1-1.) Hence,

$$A = \begin{cases} \sqrt{2S_R} & \text{Unipolar} \\ \sqrt{4S_R} & \text{Polar} \end{cases} \tag{8}$$

and the factor of $\sqrt{2}$ can make a considerable difference in the value of P_e.

Since the noise has zero mean, the variance σ^2 equals the noise power N_R at the output of the filter. Therefore, we can write $A/2\sigma$ in terms of the signal-to-noise power ratio $(S/N)_R = S_R/N_R$, namely

$$\left(\frac{A}{2\sigma}\right)^2 = \frac{A^2}{4N_R} = \begin{cases} \frac{1}{2}(S/N)_R & \text{Unipolar} \\ (S/N)_R & \text{Polar} \end{cases} \tag{9}$$

But Eq. (9) conceals the effect of the signaling rate r_b. In order to pass pulses of duration $T_b = 1/r_b$, the noise-limiting filter must have $B_N \geq r_b/2$, so

$$N_R = \eta B_N \geq \eta r_b/2 \tag{10}$$

Rapid signaling thus requires more signal power to maintain a given error probability P_e.

Example 11.2-1 Suppose a computer produces unipolar pulses at the rate $r_b = 10^6$ bps = 1 Mbps for transmission over a noisy system with $\eta = 4 \times 10^{-20}$ W/Hz. The error rate is specified to be no greater than one bit per hour, or $P_e \leq 1/3600 r_b \approx 3 \times 10^{-10}$. Table T.6 indicates that we need $A/2\sigma \geq 6.2$, and Eqs. (9) and (10) give the corresponding signal-power requirement

$$S_R = 2\left(\frac{A}{2\sigma}\right)^2 N_R \geq 1.5 \times 10^{-12} = 1.5 \text{ pW}$$

Clearly, any reasonable signal power ensures almost *errorless* transmission insofar as additive noise is concerned. Hardware glitches and other effects would be the limiting factors on system performance.

Exercise 11.2-1 Consider a unipolar system with equally likely digits and $(S/N)_R = 50$. Calculate P_{e0}, P_{e1}, and P_e when the threshold is set at the nonoptimum value $V = 0.4A$. Compare P_e with the minimum value from Eq. (7).

Regenerative Repeaters

Long-haul transmission requires repeaters, be it for analog or digital communication. But unlike analog-message repeaters, digital repeaters can be *regenerative*. 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. (11), Sect. 5.4, applies, the final signal-to-noise ratio is $(S/N)_R = (1/m)(S/N)_1$ and

$$P_e = Q\left[\sqrt{\frac{1}{m}\left(\frac{S}{N}\right)_1}\right] \tag{11}$$

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. (11) stems from the fact that the contaminating noise progressively builds up from repeater to repeater.

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 α be the error probability at each repeater, namely,

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

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* digits.) The probability of i errors in m successive conversions is given by the *binomial frequency function*, Eq. (1), Sect. 4.4,

$$P_I(i) = \binom{m}{i}\alpha^i(1-\alpha)^{m-i}$$

Since we have a destination error only when i is *odd*,

$$\begin{aligned} P_e &= \sum_{i \text{ odd}} P_I(i) \\ &= \binom{m}{1}\alpha(1-\alpha)^{m-1} + \binom{m}{3}\alpha^3(1-\alpha)^{m-3} + \cdots \\ &\approx m\alpha \end{aligned}$$

where the approximation applies for $\alpha \ll 1$ and m not too large. Hence,

$$P_e \approx mQ\left[\sqrt{\left(\frac{S}{N}\right)_1}\right] \tag{12}$$

so P_e increases linearly with m, which generally requires a much smaller power increase to counteract than Eq. (11).

Figure 11.2-5 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.

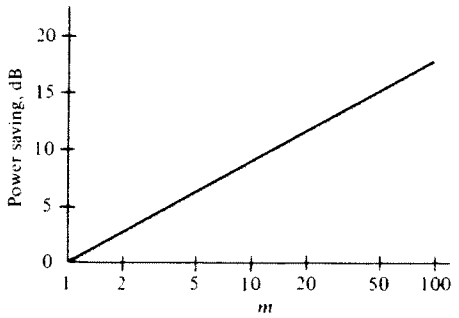


Figure 11.2-5 Power saving gained by m regenerative repeaters when $P_e = 10^{-5}$.

Matched Filtering

Every baseband digital receiver — whether at the destination or part of a regenerative repeater — should include a lowpass filter designed to remove excess noise without introducing ISI. But what's the *optimum* filter design for this purpose? For the case of timelimited pulses in white noise, the answer is a *matched filter*. We'll pursue that case here and develop the resulting minimum error probability for binary signaling in white noise.

Let the received signal be a single timelimited pulse of duration τ centered at time $t = kD$, so

$$x(t) = a_k p(t - kD)$$

where $p(0) = 1$, $p(t) = 0$ for $|t| > \tau/2$, and $\tau \leq D$. Maximizing the output ratio $(a_k/\sigma)^2$ at time $t_k = kD + t_d$ will minimize the error probability. As we learned in Sect. 5.4, this maximization calls for a matched filter whose impulse response is proportional to $p(t_d - t)$. In particular, we take

$$h(t) = \frac{1}{\tau_{eq}} p(t_d - t) \tag{13a}$$

with

$$\tau_{eq} = \int_{-\infty}^{\infty} p^2(t)\, dt \qquad t_d = \tau/2 \tag{13b}$$

The delay $t_d = \tau/2$ is the minimum value that yields a causal impulse response, and the proportionality constant $1/\tau_{eq}$ has been chosen so that the peak output amplitude equals a_k. The parameter τ_{eq} can be interpreted from the property that $a_k^2 \tau_{eq}$ equals the *energy* of the pulse $x(t)$.

In absence of noise, the resulting output pulse is $y(t) = h(t) * x(t)$, with peak value $y(t_k) = a_k$ as desired. This peak occurs $\tau/2$ seconds after the peak of $x(t)$. Thus, matched filtering introduces an unavoidable *delay*. However, it does not introduce ISI at the sampling times for adjacent pulses since $y(t) = 0$ outside of $t_k \pm \tau$. Figure 11.2-6 illustrates these points taking a rectangular shape for $p(t)$, in which case $\tau_{eq} = \tau$ and $y(t)$ has a triangular shape.

When $x(t)$ is accompanied by white noise, the output noise power from the matched filter will be

$$N_R = \sigma^2 = \frac{\eta}{2} \int_{-\infty}^{\infty} |H(f)|^2\, df$$

$$= \frac{\eta}{2} \int_{-\infty}^{\infty} |h(t)|^2\, dt = \frac{\eta}{2\tau_{eq}} \tag{14}$$

This result agrees with the lower bound in Eq. (10) since, for binary signaling, $\tau_{eq} \leq T_b = 1/r_b$. We'll use this result to evaluate the maximum value of $(A/2\sigma)^2$ and the corresponding minimum binary error probability when the noise is white and gaussian and the receiver has an optimum filter matched to the timelimited pulse shape.

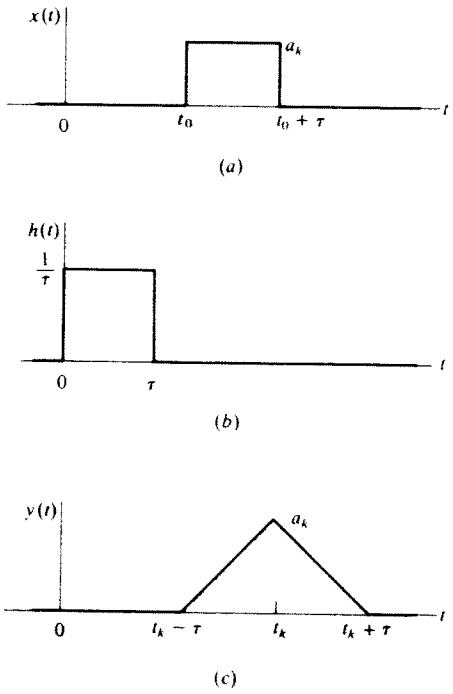


Figure 11.2-6 Matched filtering with rectangular pulses. (*a*) Received pulse; (*b*) impulse response; (*c*) output pulse.

Consider a binary transmission system with rate r_b, average received power S_R, and noise density η. We can characterize this system in terms of two new parameters E_b and γ_b defined by

$$E_b \triangleq S_R/r_b \tag{15a}$$

$$\gamma_b \triangleq S_R/\eta r_b = E_b/\eta \tag{15b}$$

The quantity E_b corresponds to the *average energy per bit*, while γ_b represents the ratio of bit energy to noise density†. If the signal consists of timelimited pulses $p(t)$ with amplitude sequence a_k, then

$$E_b = \overline{a_k^2} \int_{-\infty}^{\infty} p^2(t)\, dt = \overline{a_k^2} \tau_{eq}$$

where $\overline{a_k^2} = A^2/2$ for a unipolar signal or $\overline{a_k^2} = A^2/4$ for a polar signal. Thus, since the output noise power from a matched filter is $\sigma^2 = \eta/2\tau_{eq}$, we have

$$(A/2\sigma)^2 = \begin{cases} E_b/\eta = \gamma_b & \text{Unipolar} \\ 2E_b/\eta = 2\gamma_b & \text{Polar} \end{cases}$$

† Many authors write this ratio as E_b/N_0 where $N_0 = \eta$.

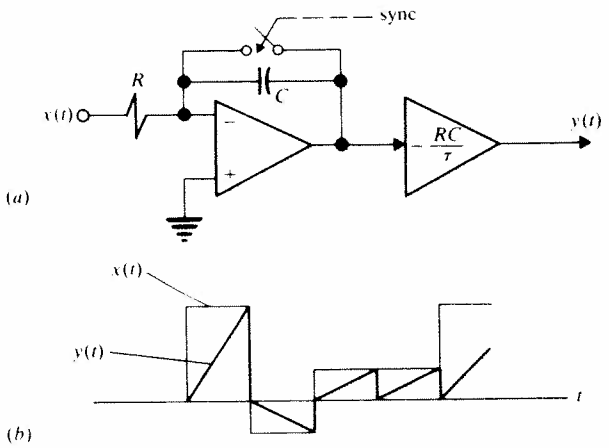


Figure 11.2-7 Integrate-and-dump filter. (*a*) op-amp circuit; (*b*) polar M-ary waveforms.

and Eq. (7) becomes

$$P_e = \begin{cases} Q(\sqrt{\gamma_b}) & \text{Unipolar} \\ Q(\sqrt{2\gamma_b}) & \text{Polar} \end{cases} \tag{16}$$

This is the minimum possible error probability, attainable only with matched filtering.

Finally, we should give some attention to the implementation of a matched filter described by Eq. (13). The impulse response for an arbitrary $p(t)$ can be approximated with passive circuit elements, but considerable design effort must be expended to get $h(t) \approx 0$ for $t > \tau$. Otherwise, the filter may produce significant ISI. For a rectangular pulse shape, you can use an active circuit such as the one diagrammed in Fig. 11.2-7*a*, called an *integrate-and-dump* filter. The op-amp integrator integrates each incoming pulse so that $y(t_k) = a_k$ at the end of the pulse, after which the dump switch resets the integrator to zero—thereby ensuring no ISI at subsequent sampling times. The integrate-and-dump filter is probably the best practical implementation of matched filtering. Figure 11.2-7*b* illustrates its operation with a polar M-ary waveform.

Exercise 11.2-2 Let $x(t)$ be the unipolar RZ signal in Fig. 11.1-1*a*. (*a*) Sketch the corresponding output waveform from a matched filter and from an integrate-and-dump filter. (*b*) Confirm that matched filtering yields $(A/2\sigma)^2 = \gamma_b$ even though $\sigma^2 = \eta r_b$ so $N_R > \eta r_b/2$.

M-ary Error Probabilities

Binary signaling provides the greatest immunity to noise for a given S/N because it has only two amplitude levels—and you can't send information with fewer

than two levels. Multilevel M-ary signaling requires more signal power but less transmission bandwidth because the signaling rate will be smaller than the bit rate of an equivalent binary signal. Consequently, M-ary signaling suits applications such as digital transmission over voice channels where the available bandwidth is limited and the signal-to-noise ratio is relatively large.

Here we calculate M-ary error probabilities in zero-mean gaussian noise. We'll take the most common case of polar signaling with an even number of equispaced levels at

$$a_k = \pm A/2, \pm 3A/2, \ldots, \pm(M-1)A/2 \tag{17}$$

We'll also assume equally likely M-ary symbols, so that

$$P_e = \frac{1}{M}(P_{e0} + P_{e1} + \cdots + P_{eM-1}) \tag{18}$$

which is the M-ary version of Eq. (5*a*).

Figure 11.2-8 shows the conditional PDFs for a quaternary ($M = 4$) polar signal plus gaussian noise. The decision rule for regeneration now involves *three* threshold levels, indicated in the figure at $y = -A$, 0, and $+A$. These are the optimum thresholds for minimizing P_e, but they do not result in equal error probabilities for all symbols. For the two extreme levels at $a_k = \pm 3A/2$ we get

$$P_{e0} = P_{e3} = Q(A/2\sigma)$$

whereas

$$P_{e1} = P_{e2} = 2Q(A/2\sigma)$$

because both positive and negative noise excursions produce errors for the inner levels at $a_k = \pm A/2$. The resulting average error probability is

$$P_e = \frac{1}{4} \times 6Q\left(\frac{A}{2\sigma}\right) = \frac{3}{2}Q\left(\frac{A}{2\sigma}\right)$$

or 50% greater than binary signaling with the same level spacing.

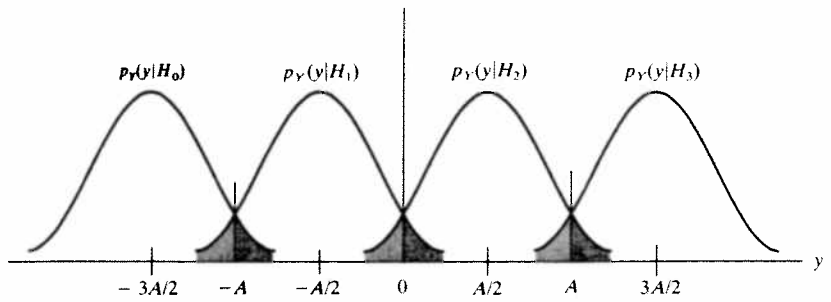


Figure 11.2-8 Conditional PDFs for a quaternary polar signal with gaussian noise.

The foregoing analysis readily generalizes to an arbitrary even value of M with $M-1$ decision thresholds at

$$y = 0, \pm A, \pm 2A, \ldots, \pm \frac{M-2}{2} A \tag{19}$$

Then $P_{e_0} = P_{e_{M-1}} = Q(A/2\sigma)$ while the $M-2$ inner levels have doubled error probability, yielding the average error probability

$$\begin{aligned} P_e &= \frac{1}{M}\left[2 \times Q\left(\frac{A}{2\sigma}\right) + (M-2) \times 2Q\left(\frac{A}{2\sigma}\right)\right] \\ &= \frac{2M-2}{M} Q\left(\frac{A}{2\sigma}\right) = 2\left(1 - \frac{1}{M}\right) Q\left(\frac{A}{2\sigma}\right) \end{aligned} \tag{20}$$

Equation (20) clearly reduces to Eq. (7) when $M = 2$, whereas $P_e \approx 2Q(A/2\sigma)$ when $M \gg 2$.

Next, we relate $A/2\sigma$ to the signal power and noise density assuming a time-limited pulse shape $p(t)$ so the average energy per M-ary digit is $E_M = \overline{a_k^2}\tau_{eq}$ where

$$\tau_{eq} = \int_{-\infty}^{\infty} p^2(t)\, dt$$

as before. If the M amplitude levels are equally likely and given by Eq. (17), then

$$\overline{a_k^2} = 2 \times \frac{1}{M} \sum_{i=1}^{M/2} (2i-1)^2 \left(\frac{A}{2}\right)^2 = \frac{M^2-1}{12} A^2 \tag{21}$$

Hence, since $S_R = rE_M$,

$$\begin{aligned} \left(\frac{A}{2\sigma}\right)^2 &= \frac{3}{M^2-1} \frac{E_M}{\tau_{eq}\sigma^2} \\ &= \frac{3}{M^2-1} \frac{1}{r\tau_{eq}} \frac{S_R}{N_R} \le \frac{6}{M^2-1} \frac{S_R}{\eta r} \end{aligned} \tag{22}$$

where the upper bound corresponds to $N_R = \eta/2\tau_{eq}$ obtained with matched filtering. Equations (20) and (22) constitute our final result for error probability in a polar M-ary system with gaussian white noise.

More often than not, M-ary signaling is used to transmit *binary* messages and the value of M is selected by the system engineer to best fit the available channel. We should therefore investigate the design considerations in the selection of M, especially the impact on error probability. But Eqs. (20) and (22) fail to tell the full story for two reasons: first, the M-ary signaling rate differs from the bit rate r_b; second, the M-ary error probability differs from the bit error probability.

We can easily account for the signaling-rate difference when the message bits are encoded in blocks of length $\log_2 M$. Then r_b and r are related by

$$r_b = r \log_2 M \tag{23}$$

from Eq. (4), Sect. 11.1. To relate the M-ary symbol error probability P_e to the resulting error probability per bit, we'll assume a *Gray code* and a reasonably large signal-to-noise ratio. Under these conditions a noise excursion seldom goes beyond one amplitude level in the M-ary waveform, which corresponds to just one erroneous bit in the block of $\log_2 M$ bits. Therefore,

$$P_{be} \approx P_e/\log_2 M \tag{24}$$

where P_{be} stands for the equivalent *bit error probability*, also called the bit error rate (BER).

Combining Eqs. (23) and (24) with our previous M-ary expressions, we finally have

$$P_{be} \approx 2 \frac{M-1}{M \log_2 M} Q\left(\frac{A}{2\sigma}\right) \tag{25a}$$

in which

$$\left(\frac{A}{2\sigma}\right)^2 \le \frac{6}{M^2-1} \frac{S_R}{\eta r} = \frac{6 \log_2 M}{M^2-1} \gamma_b \tag{25b}$$

Notice that the upper bound with matched filtering has been written in terms of $\gamma_b = S_R/\eta r_b = S_R/(\eta r \log_2 M)$. This facilitates the study of M-ary signaling as a function of energy per message bit.

Example 11.2-2 Comparison of binary and M-ary signaling Suppose the channel in question has a fixed signaling rate $r = 3000$ baud $= 3$ kbaud and a fixed signal-to-noise ratio $(S/N)_R = 400 \approx 26$ dB. (These values would be typical of a voice telephone channel, for instance.) We'll assume matched filtering of NRZ rectangular pulses, so $r\tau_{eq} = 1$ and

$$\left(\frac{A}{2\sigma}\right)^2 = \frac{3}{M^2-1} (S/N)_R = \frac{6 \log_2 M}{M^2-1} \gamma_b$$

which follow from Eqs. (22) and (25*b*).

Binary signaling yields a vanishingly small error probability when $(S/N)_R = 400$, but at the rather slow rate $r_b = r = 3$ kbps. M-ary signaling increases the bit rate, per Eq. (23), but the error probability also increases because the spacing between amplitude levels gets smaller when you increase M with the signal power held fixed. Table 11.2-1 brings out the trade-off between bit rate and error probability for this channel.

Table 11.2-1 M-ary signaling with $r = 3$ kilobaud and $(S/N)_R = 400$

M	r_b(kbps)	$A/2\sigma$	P_{be}
2	3	20.0	3×10^{-89}
4	6	8.9	1×10^{-19}
8	9	4.4	4×10^{-6}
16	12	2.2	7×10^{-3}
32	15	1.1	6×10^{-2}

Table 11.2-2 M-ary signaling with $r_b = 9$ kbps and $P_{be} = 4 \times 10^{-6}$

M	r (kbaud)	γ_b
2	9.00	10
4	4.50	24
8	3.00	67
16	2.25	200
32	1.80	620

Another type of trade-off is illustrated by Table 11.2-2, where the bit rate and error probability are both held fixed. Increasing M then yields a lower signaling rate r—implying a smaller transmission bandwidth requirement. However, the energy per bit must now be increased to keep the error probability unchanged. Observe that going from $M = 2$ to $M = 32$ reduces r by 1/5 but increases γ_b by more than a factor of 60. This type of trade-off will be reconsidered from the broader viewpoint of information theory in Chap. 15.

Exercise 11.2-3 Consider the three-level bipolar binary format in Fig. 11.1-1*c* with amplitude probabilities $P(a_k = 0) = 1/2$ and $P(a_k = +A) = P(a_k = -A) = 1/4$. Make a sketch similar to Fig. 11.2-8 and find P_e in terms of A and σ when the decision thresholds are at $y = \pm A/2$. Then calculate S_R and express P_e in a form like Eq. (16).

11.3 BANDLIMITED DIGITAL PAM SYSTEMS

This section develops design procedures for baseband digital systems when the transmission channel imposes a bandwidth limitation. By this we mean that the available transmission bandwidth is not large compared to the desired signaling rate and, consequently, rectangular signaling pulses would be severely distorted. Instead, we must use *bandlimited* pulses specially shaped to avoid ISI.

Accordingly, we begin with Nyquist's strategy for bandlimited pulse shaping. Then we consider the optimum terminal filters needed to minimize error probability. The assumption that the noise has a gaussian distribution with zero mean value will be continued, but we'll allow an arbitrary noise power spectrum. We'll also make allowance for linear transmission distortion, which leads to the subject of equalization for digital systems. The section closes with an introductory discussion of correlative coding techniques that increase the signaling rate on a bandlimited channel.

Nyquist Pulse Shaping

Our presentation of Nyquist pulse shaping will be couched in general terms of M-ary signaling with $M \geq 2$ and symbol interval $D = 1/r$. In order to focus on

potential ISI problems at the receiver, we'll let $p(t)$ be the pulse shape at the output of the receiving filter. Again assuming that the transmitter gain compensates for transmission loss, the output waveform in absence of noise is

$$y(t) = \sum_k a_k p(t - t_d - kD)$$

As before, we want $p(t)$ to have the property

$$p(t) = \begin{cases} 1 & t = 0 \\ 0 & t = \pm D, \pm 2D, \ldots \end{cases} \tag{1a}$$

which eliminates ISI. Now we impose the additional requirement that the pulse spectrum be *bandlimited*, such that

$$P(f) = 0 \qquad |f| \geq B \tag{1b}$$

where

$$B = \frac{r}{2} + \beta \qquad 0 \leq \beta \leq \frac{r}{2}$$

This spectral requirement permits signaling at the rate

$$r = 2(B - \beta) \qquad B \leq r \leq 2B \tag{2}$$

in which B may be interpreted as the minimum required transmission bandwidth, so that $B_T \geq B$.

Nyquist's *vestigial-symmetry theorem* states that Eq. (1) is satisfied if $p(t)$ has the form

$$p(t) = p_\beta(t) \operatorname{sinc} rt \tag{3a}$$

with

$$\begin{aligned} \mathcal{F}[p_\beta(t)] = P_\beta(f) = 0 \qquad |f| > \beta \\ p_\beta(0) = \int_{-\infty}^{\infty} P_\beta(f)\, df = 1 \end{aligned} \tag{3b}$$

Clearly, $p(t)$ has the time-domain properties of Eq. (1*a*). It also has the frequency-domain properties of Eq. (1*b*) since

$$P(f) = P_\beta(f) * [(1/r)\Pi(f/r)]$$

and the convolution of two bandlimited spectra produces a new bandlimited spectrum whose bandwidth equals the sum of the bandwidths, namely, $B = \beta + r/2$. Usually we take $p_\beta(t)$ to be an even function so $P_\beta(f)$ is real and even; then $P(f)$ has vestigial symmetry around $f = \pm r/2$, like the symmetry of a vestigial sideband filter.

Infinitely many functions satisfy Nyquist's conditions, including the case when $p_\beta(t) = 1$ so $\beta = 0$ and $p(t) = \operatorname{sinc} rt$, as in Eq. (6), Sect. 11.1. We know that this pulse shape allows bandlimited signaling at the maximum rate $r = 2B$.

However, synchronization turns out to be a very touchy matter because the pulse shape falls off no faster than $1/|t|$ as $|t| \to \infty$. Consequently, a small timing error ϵ results in the sample value

$$y(t_K) = a_K \operatorname{sinc} r\epsilon + \sum_{k \neq K} a_k \operatorname{sinc} (KD - kD + r\epsilon)$$

and the ISI in the second term can be quite large.

Synchronization problems are eased by reducing the signaling rate and using pulses with a *cosine rolloff* spectrum. Specifically, if

$$P_\beta(f) = \frac{\pi}{4\beta} \cos \frac{\pi f}{2\beta} \Pi\left(\frac{f}{2\beta}\right) \tag{4a}$$

then

$$P(f) = \begin{cases} \dfrac{1}{r} & |f| < \dfrac{r}{2} - \beta \\ \dfrac{1}{r} \cos^2 \dfrac{\pi}{4\beta}\left(|f| - \dfrac{r}{2} + \beta\right) & \dfrac{r}{2} - \beta < |f| < \dfrac{r}{2} + \beta \\ 0 & |f| > \dfrac{r}{2} + \beta \end{cases} \tag{4b}$$

and the corresponding pulse shape is

$$p(t) = \frac{\cos 2\pi\beta t}{1 - (4\beta t)^2} \operatorname{sinc} rt \tag{5}$$

Plots of $P(f)$ and $p(t)$ are shown in Fig. 11.3-1 for two values of β along with $\beta = 0$. When $\beta > 0$, the spectrum has a smooth rolloff and the leading and trailing oscillations of $p(t)$ decay more rapidly than those of sinc rt.

Further consideration of Eqs. (4) and (5) reveals two other helpful properties of $p(t)$ in the special case when $\beta = r/2$, known as 100% rolloff. The spectrum

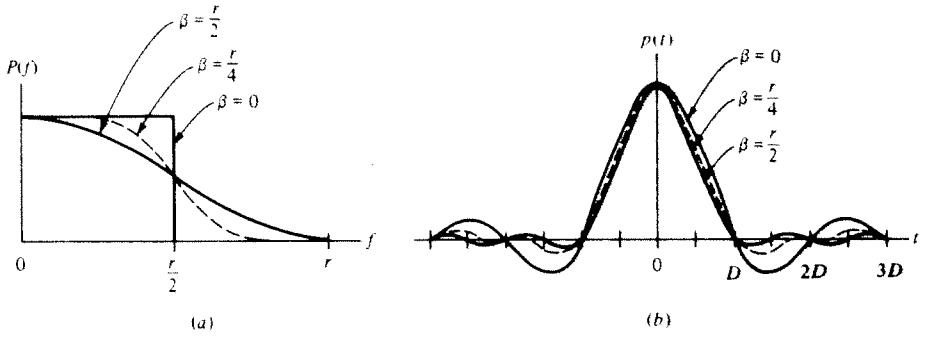


Figure 11.3-1 Nyquist pulse shaping. (*a*) Spectra; (*b*) waveforms.

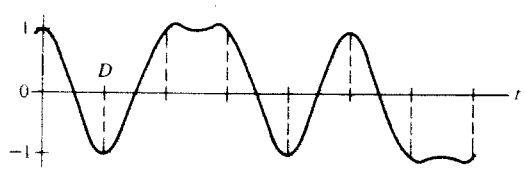


Figure 11.3-2 Baseband waveform for 10110100 using Nyquist pulses with $\beta = r/2$.

then reduces to the *raised cosine* shape

$$P(f) = \frac{1}{r} \cos^2 \frac{\pi f}{2r} = \frac{1}{2r}\left[1 + \cos\left(\frac{\pi f}{r}\right)\right] \qquad |f| \leq r \tag{6a}$$

and

$$p(t) = \frac{\operatorname{sinc} 2rt}{1 - (2rt)^2} \tag{6b}$$

The half-amplitude width of this pulse exactly equals the symbol interval D, that is, $p(\pm 0.5D) = 1/2$, and there are additional zero crossings at $t = \pm 1.5D$, $\pm 2.5D, \ldots$. A polar signal constructed with this pulse shape will therefore have zero crossings precisely halfway between the pulse centers whenever the amplitude changes polarity. Figure 11.3-2 illustrates this feature with the binary message 10110100. These zero crossings make it a simple task to extract a synchronization signal for timing purposes at the receiver. However, the penalty is a 50% reduction of signaling speed since $r = B$ rather than $2B$. Nyquist proved that the pulse shape defined by Eq. (6) is the *only* one possessing all of the aforementioned properties.

Exercise 11.3-1 Sketch $P(f)$ and find $p(t)$ for the Nyquist pulse generated by taking $P_\beta(f) = (2/r)\Lambda(2f/r)$. Compare your results with Eq. (6).

Optimum Terminal Filters

Having abandoned rectangular pulses, we must likewise abandon the conventional matched filter and reconsider the design of the optimum receiving filter that minimizes error probability. This turns out to be a relatively straightforward problem under the following reasonable conditions:

1. The signal format is polar, and the amplitudes a_k are uncorrelated and equally likely.
2. The transmission channel is linear but not necessarily distortionless.
3. The filtered output pulse $p(t)$ is to be Nyquist shaped.
4. The noise is additive and has a zero-mean gaussian distribution but may have a nonwhite power spectrum.

To allow for possible channel distortion and/or nonwhite noise, our optimization must involve filters at both the transmitter and receiver. As a bonus, the source waveform $x(t)$ may have a more-or-less arbitrary pulse shape $p_x(t)$.

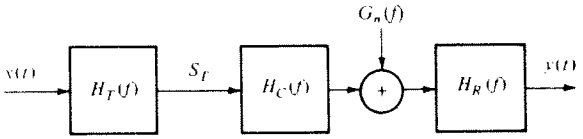


Figure 11.3-3

Figure 11.3-3 lays out the system diagram, including a transmitting filter function $H_T(f)$, a channel function $H_C(f)$, and a receiving filter function $H_R(f)$. The input signal has the form

$$x(t) = \sum_k a_k p_x(t - kD) \tag{7a}$$

and its power spectrum is

$$G_x(f) = \sigma_a^2 r |P_x(f)|^2 \tag{7b}$$

where $P_x(f) = \mathcal{F}[p_x(t)]$ and

$$\sigma_a^2 = \overline{a_k^2} = \frac{M^2 - 1}{12} A^2 \tag{7c}$$

These relations follow from Eq. (12), Sect. 11.1, and Eq. (21), Sect. 11.2, with our stated conditions on a_k. Thus, the transmitted signal power will be

$$\begin{aligned} S_T &= \int_{-\infty}^{\infty} |H_T(f)|^2 G_x(f)\, df \\ &= \frac{M^2 - 1}{12} A^2 r \int_{-\infty}^{\infty} |H_T(f) P_x(f)|^2\, df \end{aligned} \tag{8}$$

a result we'll need shortly.

At the output of the receiving filter we want the input pulse $p_x(t)$ to produce a Nyquist-shaped pulse $p(t - t_d)$, where t_d represents any transmission time delay. The transfer functions in Fig. 11.3-3 must therefore obey the relationship

$$P_x(f) H_T(f) H_C(f) H_R(f) = P(f) e^{-j\omega t_d} \tag{9}$$

so both terminal filters help shape $p(t)$. But only the receiving filter controls the output noise power

$$N_R = \sigma^2 = \int_{-\infty}^{\infty} |H_R(f)|^2 G_n(f)\, df \tag{10}$$

where $G_n(f)$ is the noise power spectrum at the input to the receiver.

Equations (7)–(10) constitute the information relevant to our design problem. Specifically, since the error probability decreases as $A/2\sigma$ increases, we seek the terminal filters that maximize $(A/2\sigma)^2$ subject to two constraints: (1) the transmitted power must be held fixed at some specified value S_T, and (2) the filter transfer functions must satisfy Eq. (9).

We incorporate Eq. (9) and temporarily eliminate $H_T(f)$ by writing

$$|H_T(f)| = \frac{|P(f)|}{|P_x(f) H_C(f) H_R(f)|} \tag{11}$$

Then we use Eqs. (8) and (10) to express $(A/2\sigma)^2$ as

$$\left(\frac{A}{2\sigma}\right)^2 = \frac{3S_T}{(M^2 - 1)r} \frac{1}{I_{HR}} \tag{12a}$$

where

$$I_{HR} = \int_{-\infty}^{\infty} |H_R(f)|^2 G_n(f)\, df \int_{-\infty}^{\infty} \frac{|P(f)|^2}{|H_C(f) H_R(f)|^2}\, df \tag{12b}$$

Maximizing $(A/2\sigma)^2$ thus boils down to minimizing the product of integrals I_{HR}, in which $H_R(f)$ is the only function under our control.

Now observe that Eq. (12b) has the form of the right-hand side of Schwarz's inequality as stated in Eq. (17), Sect. 3.5. The minimum value of I_{HR} therefore occurs when the two integrands are proportional. Consequently, the optimum receiving filter has

$$|H_R(f)|^2 = \frac{g|P(f)|}{\sqrt{G_n(f)}\,|H_C(f)|} \tag{13a}$$

where g is an arbitrary gain constant. Equation (11) then gives the optimum transmitting filter characteristic

$$|H_T(f)|^2 = \frac{|P(f)|\sqrt{G_n(f)}}{g|P_x(f)|^2 |H_C(f)|} \tag{13b}$$

These expressions specify the optimum amplitude ratios for the terminal filters. Note that the receiving filter deemphasizes those frequencies where $G_n(f)$ is large, and the transmitting filter supplies the corresponding preemphasis. The phase shifts are arbitrary, providing that they satisfy Eq. (9).

Substituting Eq. (13) into Eq. (12) yields our final result

$$\left(\frac{A}{2\sigma}\right)_{\max}^2 = \frac{3S_T}{(M^2 - 1)r} \left[\int_{-\infty}^{\infty} \frac{|P(f)|\sqrt{G_n(f)}}{|H_C(f)|}\, df\right]^{-2} \tag{14}$$

from which the error probability can be calculated using Eq. (20), Sect. 11.2. As a check of Eq. (14), take the case of white noise with $G_n(f) = \eta/2$ and a distortionless channel with transmission loss L so $|H_C(f)|^2 = 1/L$; then

$$\left(\frac{A}{2\sigma}\right)_{\max}^2 = \frac{6S_T/L}{(M^2 - 1)r\eta} \left[\int_{-\infty}^{\infty} |P(f)|\, df\right]^{-2}$$

But since $S_T/L = S_R$ and Nyquist-shaped pulses have

$$\int_{-\infty}^{\infty} |P(f)|\, df = 1$$

we thus obtain

$$\left(\frac{A}{2\sigma}\right)_{\max}^2 = \frac{6}{M^2 - 1} \frac{S_R}{\eta r} = \frac{6 \log_2 M}{M^2 - 1} \gamma_b$$

which confirms that the optimum terminal filters yield the same upper bound as matched filtering — see Eqs. (22) and (25), Sect. 11.2.

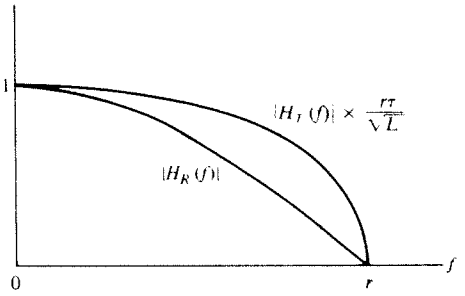


Figure 11.3-4 Amplitude ratio of optimum filters in Example 11.3-1.

Example 11.3-1 Consider a system with white noise, transmission loss L, and a distortionless channel response over $|f| \le B_T$ where $B_T \ge r$. This transmission bandwidth allows us to use the pulse shape $p(t)$ in Eq. (6), thereby simplifying synchronization. Simplicity also suggests using the rectangular input pulse $p_x(t) = \Pi(t/\tau)$ with $\tau \le 1/r$, so $P_x(f) = \tau \operatorname{sinc} f\tau$. Taking the gain constant g in Eq. (13) such that $|H_R(0)| = 1$, we have

$$|H_R(f)| = \cos\frac{\pi f}{2r} \qquad |H_T(f)| = \sqrt{L}\,\frac{\cos(\pi f/2r)}{r\tau \operatorname{sinc} f\tau} \qquad |f| \le r$$

as plotted in Fig. 11.3-4. Notice the slight high-frequency rise in $|H_T(f)|$ compared to $|H_R(f)|$. If the input pulses have a small duration $\tau \ll 1/r$, then this rise becomes negligible and $|H_T(f)| \approx |H_R(f)|$, so one circuit design serves for both filters.

Exercise 11.3-2 Carry out the details going from Eq. (12) to Eqs. (13) and (14).

Equalization

Regardless of which particular pulse shape has been chosen, some amount of residual ISI inevitably remains in the output signal as a result of imperfect filter design, incomplete knowledge of the channel characteristics, etc. Hence, an adjustable equalizing filter is often inserted between the receiving filter and the regenerator. These "mop-up" equalizers usually have the structure of a transversal filter previously discussed in Sect. 3.2 relative to linear distortion of analog signals. However, mop-up equalization of digital signals involves different design strategies that deserve attention here.

Figure 11.3-5 shows a transversal equalizer with $2N + 1$ taps and total delay $2ND$. The distorted pulse shape $\tilde{p}(t)$ at the input to the equalizer is assumed to have its peak at $t = 0$ and ISI on both sides. The equalized output pulse will be

$$p_{eq}(t) = \sum_{n=-N}^{N} c_n \tilde{p}(t - nD - ND)$$

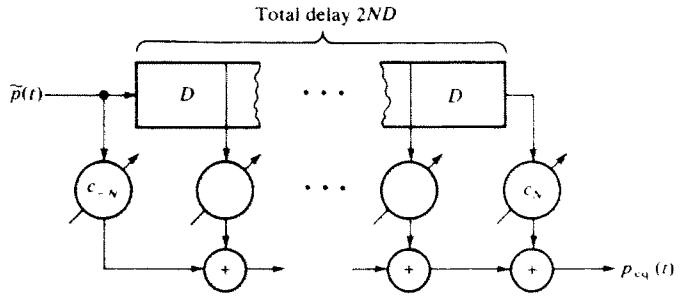


Figure 11.3-5 Transversal equalizer with $2N + 1$ taps.

and sampling at $t_k = kD + ND$ yields

$$p_{eq}(t_k) = \sum_{n=-N}^{N} c_n \tilde{p}(kD - nD) = \sum_{n=-N}^{N} c_n \tilde{p}_{k-n} \tag{15}$$

where we've introduced the shorthand notation $\tilde{p}_{k-n} = \tilde{p}[(k-n)D]$. Equation (15) thus takes the form of a *discrete convolution.*

Ideally, we would like the equalizer to eliminate all ISI, resulting in

$$p_{eq}(t_k) = \begin{cases} 1 & k = 0 \\ 0 & k \ne 0 \end{cases}$$

But this cannot be achieved, in general, because the $2N + 1$ tap gains are the only variables at our disposal. We might settle instead for choosing the tap gains such that

$$p_{eq}(t_k) = \begin{cases} 1 & k = 0 \\ 0 & k = \pm 1, \pm 2, \ldots, \pm N \end{cases} \tag{16}$$

thereby forcing N zero values on each side of the peak of $p_{eq}(t)$. The corresponding tap gains are computed from Eqs. (15) and (16) combined in the matrix equation

$$\begin{bmatrix} \tilde{p}_0 & \cdots & \tilde{p}_{-2N} \\ \vdots & & \vdots \\ \tilde{p}_{N-1} & \cdots & \tilde{p}_{-N-1} \\ \tilde{p}_N & \cdots & \tilde{p}_{-N} \\ \tilde{p}_{N+1} & \cdots & \tilde{p}_{-N+1} \\ \vdots & & \vdots \\ \tilde{p}_{2N} & \cdots & \tilde{p}_0 \end{bmatrix} \begin{bmatrix} c_{-N} \\ \vdots \\ c_{-1} \\ c_0 \\ c_1 \\ \vdots \\ c_N \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{17}$$

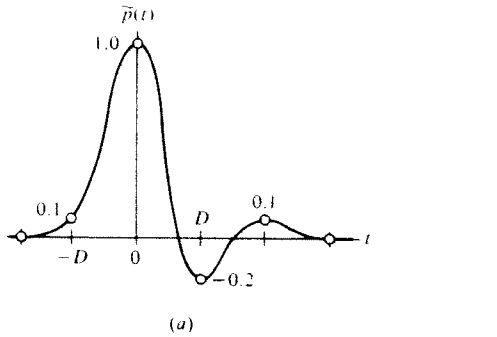
Equation (17) describes a *zero-forcing equalizer.* This equalization strategy is optimum in the sense that it minimizes the peak intersymbol interference, and it has the added advantage of simplicity. Other optimization criteria lead to different strategies with more complicated tap-gain relationships.

When the transmission channel is a switched telephone link or a radio link with slowly changing conditions, the values of $\tilde{p}_k$ will not be known in advance. The tap gains must then be adjusted on-line using a training sequence transmitted before the actual message sequence. An *adaptive equalizer* incorporates a microprocessor for automatic rather than manual tap-gain adjustment. More sophisticated versions adjust themselves continuously using error measures derived from the message sequence. Further details regarding adaptive equalization are given by Haykin (1983, Sect. 9.6) and Proakis (1983, Chap. 6).

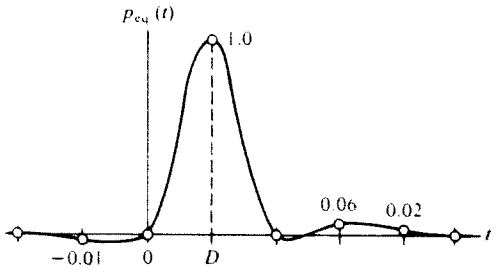
Mop-up equalization, whether fixed or adaptive, does have one hidden catch in that the equalizer somewhat increases the noise power at the input to the regenerator. This increase is usually more than compensated for by the reduction of ISI.

Example 11.3-2 Suppose a three-tap zero forcing equalizer is to be designed for the distorted pulse plotted in Fig. 11.3-6*a*. Inserting the values of $\tilde{p}_k$ into Eq. (17) with $N = 1$, we have

$$\begin{bmatrix} 1.0 & 0.1 & 0.0 \\ -0.2 & 1.0 & 0.1 \\ 0.1 & -0.2 & 1.0 \end{bmatrix} \begin{bmatrix} c_{-1} \\ c_0 \\ c_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$



(*a*)



(*b*)

Figure 11.3-6 (*a*) Distorted pulse; (*b*) equalized pulse.

Therefore,

$$c_{-1} = -0.096 \qquad c_0 = 0.96 \qquad c_1 = 0.2$$

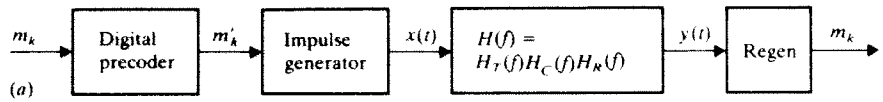
and the corresponding sample values of $p_{eq}(t)$ are plotted in Fig. 11.3-6*b* with an interpolated curve. As expected, there is one zero on each side of the peak. However, zero forcing has produced some small ISI at points further out where the unequalized pulse was zero.

Correlative Coding★

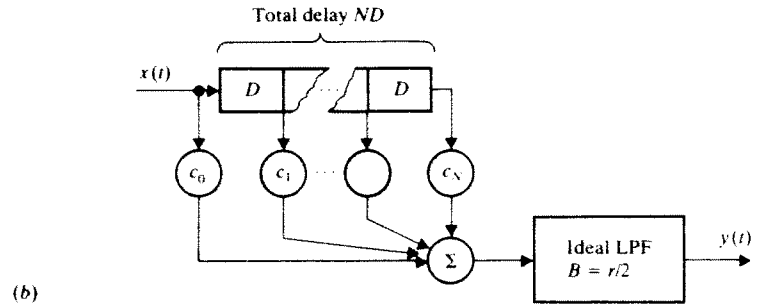
Correlative coding, also known as *partial-response signaling*, is a strategy for bandlimited transmission at $r = 2B$ that avoids the problems associated with $p(t) = \text{sinc } rt$. The strategy involves two key functional operations, correlative filtering and digital precoding. *Correlative filtering* purposely introduces controlled intersymbol interference, resulting in a pulse train with an increased number of amplitude levels and a correlated amplitude sequence. Nyquist's signaling-rate limitation no longer applies because the correlated symbols are not independent, and therefore higher signaling rates are possible. *Digital precoding* of the message sequence before waveform generation facilitates message recovery from the correlative-filtered pulse train.

Figure 11.3-7*a* shows the general model of a transmission system with correlative coding, omitting noise. The digital precoder takes the sequence of message symbols m_k and produces a coded sequence m'_k applied to an impulse generator. (In practice, the impulses would be replaced by short rectangular pulses with duration $\tau \ll D$.) The resulting input signal is an impulse train

$$x(t) = \sum_k a_k \, \delta(t - kD)$$



(*a*)



(*b*)

Figure 11.3-7 (*a*) Transmission system with correlative coding; (*b*) equivalent correlative filter.

whose weights a_k represent the m'_k. The terminal filters and channel have the overall transfer function $H(f)$, producing the output waveform

$$y(t) = \sum_k a_k h(t - kD) \tag{18}$$

where $h(t) = \mathcal{F}^{-1}[H(f)]$.

Although a correlative filter does not appear as a distinct block in Fig. 11.3-7*a*, the transfer function $H(f)$ must be equivalent to that of Fig. 11.3-7*b*—which consists of a transversal filter and an ideal LPF. The transversal filter has total delay ND and $N + 1$ tap gains. Since the impulse response of the LPF is sinc rt and $r = 1/D$, the cascade combination yields the overall impulse response

$$h(t) = \sum_{n=0}^{N} c_n \operatorname{sinc}(rt - n) \tag{19}$$

Hence, Eq. (18) becomes

$$\begin{aligned} y(t) &= \sum_k a_k \left[\sum_{n=0}^{N} c_n \operatorname{sinc}(rt - n - k) \right] \\ &= \sum_k a'_k \operatorname{sinc}(rt - k) \end{aligned} \tag{20a}$$

where

$$a'_k \triangleq c_0 a_k + c_1 a_{k-1} + \cdots + c_N a_{k-N} = \sum_{n=0}^{N} c_n a_{k-n} \tag{20b}$$

Message regeneration must then be based on the sample values $y(t_k) = a'_k$.

Equation (20) brings out the fact that correlative filtering changes the amplitude sequence a_k into the modified sequence a'_k. We say that this sequence has a *correlation span* of N symbols, since each a'_k depends on the previous N values of a_k. Furthermore, when the a_k sequence has M levels, the a'_k sequence has $M' > M$ levels. To demonstrate that these properties of correlative filtering lead to practical bandlimited transmission at $r = 2B$, we must look at a specific case.

Duobinary signaling is the simplest type of correlative coding, having $M = 2$, $N = 1$, and $c_0 = c_1 = 1$. The equivalent correlative filter is diagrammed in Fig. 11.3-8 along with its impulse response

$$h(t) = \operatorname{sinc} r_b t + \operatorname{sinc}(r_b t - 1) \tag{21a}$$

and the magnitude of the transfer function

$$H(f) = \frac{2}{r_b} \cos \frac{\pi f}{r_b} e^{-j\pi f/r_b} \qquad |f| \le r_b/2 \tag{21b}$$

The smooth rolloff of $H(f)$ is similar to the spectrum of a Nyquist-shaped pulse and can be synthesized to a good approximation. But, unlike Nyquist pulse shaping, duobinary signaling achieves this rolloff without increasing the bandwidth requirement above $B = r_b/2$. In exchange for the signaling rate advantage, a duobinary waveform has intentional ISI and $M' = 3$ levels—attributed to the property of the impulse response that $h(t) = 1$ at both $t = 0$ and $t = T_b$.

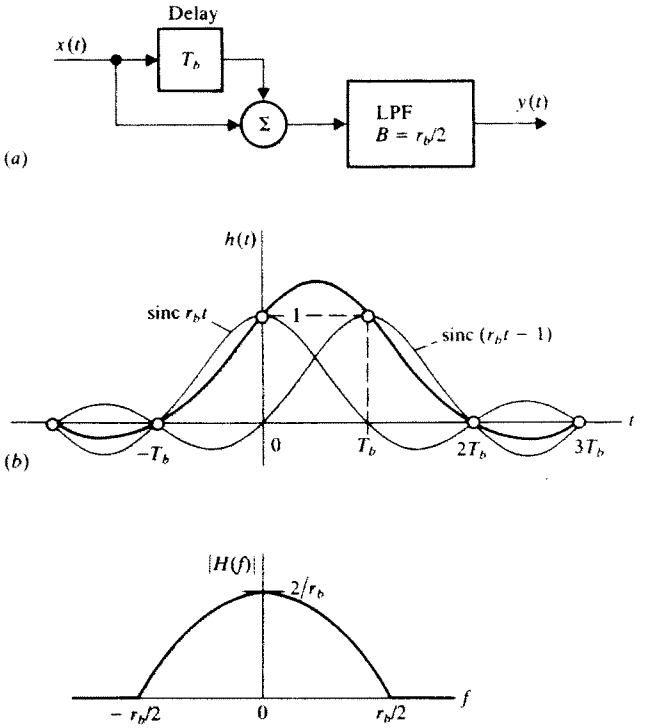


Figure 11.3-8 Duobinary signaling. (*a*) Equivalent correlative filter; (*b*) impulse response; (*c*) amplitude ratio.

To bring out the ISI effect, let the amplitude sequence a_k be related to the precoded binary message sequence m'_k by

$$a_k = (m'_k - 1/2)A = \begin{cases} +A/2 & m'_k = 1 \\ -A/2 & m'_k = 0 \end{cases} \tag{22}$$

which is equivalent to a polar binary format with level spacing A. Equation (20) then gives the corresponding output levels

$$y(t_k) = a'_k = a_k + a_{k-1} = (m'_k + m'_{k-1} - 1)A \tag{23a}$$

$$= \begin{cases} +A & m'_k = m'_{k-1} = 1 \\ 0 & m'_k \neq m'_{k-1} \\ -A & m'_k = m'_{k-1} = 0 \end{cases} \tag{23b}$$

In principle, you could use Eq. (23*b*) to recover m'_k from $y(t_k)$ if you have previously recovered m'_{k-1}. However, when noise causes an erroneous value of m'_{k-1}, all subsequent message bits will be in error until the next noise-induced error—a phenomenon called *error propagation*.

The digital precoder for duobinary signaling shown in Fig. 11.3-9*a* prevents error propagation and makes it possible to recover the input message sequence m_k from $y(t_k)$. The precoder consists of an exclusive-OR gate with feedback through a *D*-type flip-flop. Figure 11.3-9*b* lists the coding truth table along with the algebraic sum $m'_k + m'_{k-1}$ that appears in Eq. (23*a*). Substitution now yields

$$y(t_k) = \begin{cases} \pm A & m_k = 0 \\ 0 & m_k = 1 \end{cases} \tag{24}$$

which does not involve m_{k-1} thanks to the precoder. When $y(t)$ includes additive gaussian white noise, the appropriate decision rule for message regeneration is:

$$\text{Choose } m_k = 0 \quad \text{if} \quad |y| > A/2$$

$$\text{Choose } m_k = 1 \quad \text{if} \quad |y| < A/2$$

This rule is easily implemented with a rectifier and a single decision threshold set at $A/2$. Optimum terminal filter design gives the minimum error probability

$$P_e = \frac{3}{2} Q\left(\frac{\pi}{4}\sqrt{2\gamma_b}\right) \tag{25}$$

which is somewhat higher than that of a polar binary system.

When the transmission channel has poor DC response, *modified duobinary signaling* may be employed. The correlative filter has $N = 2$, $c_0 = 1$, $c_1 = 0$, and $c_2 = -1$, so that

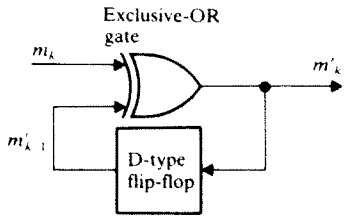
$$H(f) = \frac{2j}{r_b} \sin \frac{2\pi f}{r_b} e^{-j2\pi f/r_b} \qquad |f| \le r_b/2 \tag{26}$$

Figure 11.3-10 shows $|H(f)|$ and the block diagram of the correlative filter. The precoder takes the form of Fig. 11.3-9 with two flip-flops in series to feed back m'_{k-2}. If the pulse generation is in accordance with Eq. (22), then

$$y(t_k) = a_k - a_{k-2} = (m'_k - m'_{k-2})A$$

$$= \begin{cases} 0 & m_k = 0 \\ \pm A & m_k = 1 \end{cases} \tag{27}$$

which can be compared with Eqs. (23) and (24).



(*a*)

m_k	m'_{k-1}	m'_k	$m'_k + m'_{k-1}$
0	0	0	0
0	1	1	2
1	0	1	1
1	1	0	1

(*b*)

Figure 11.3-9 (*a*) Digital precoder for duobinary signaling; (*b*) truth table.

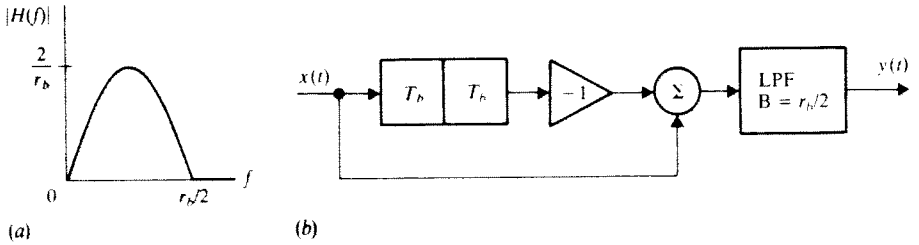


Figure 11.3-10 Correlative filter for modified duobinary signaling. (*a*) Amplitude ratio of correlative filter; (*b*) block diagram.

Exercise 11.3-3 Construct a truth table like Fig. 11.3-9*b* for the case of modified duobinary signaling and use it to obtain Eq. (27).

11.4 SYNCHRONIZATION TECHNIQUES

Synchronization is the art of making clocks tick together. The clocks in a digital communication system are at the transmitter and receiver, and allowance must be made for the transmission time delay between them. Besides *symbol* synchronization, most systems also require *frame* synchronization to identify the start of a message or various subdivisions within the message sequence. Additionally, carrier synchronization is essential for digital transmission with coherent carrier modulation—a topic to be discussed in Chap. 14.

Here we'll consider symbol and frame synchronization in baseband binary systems. Our attention will be focused on extracting synchronization from the received signal itself, rather than using an auxiliary sync signal. By way of an overview, Fig. 11.4-1 illustrates the position of the *bit synchronizer* relative to the clock and regenerator. Framing information is usually derived from the regenerated message and the clock, as indicated by the location of the *frame synchronizer*. We'll look at typical synchronization techniques, along with the related topics of shift-register operations for message scrambling and pseudonoise (PN) sequences for framing purposes.

Our coverage of synchronization will be primarily descriptive and illustrative. Detailed treatments of digital synchronization with additive noise are given by Stiffler (1971) and Gardner and Lindsey (1980).

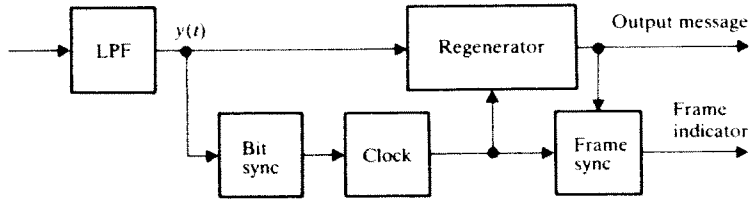


Figure 11.4-1 Synchronization in a binary receiver.

Bit Synchronization

Bit-sync generation becomes almost trivial when $y(t)$ has unipolar RZ format so its power spectrum $G_y(f)$ includes $\delta(f \pm r_b)$, like Fig. 11.1-5. A PLL or narrow BPF tuned to $f = r_b$ will then extract a sinusoid proportional to $\cos(2\pi r_b t + \phi)$, and phase adjustment yields a sync signal for the clock. The same technique works with a *polar* format if $y(t)$ is first processed by a square-law device, as diagrammed in Fig. 11.4-2*a*. The resulting unipolar waveform $y^2(t)$ shown in Fig. 11.4-2*b* now has the desired sinusoidal component at $f = r_b$. Various other nonlinear polar-to-unipolar operations on $y(t)$ achieve like results in open-loop bit synchronizers.

But a closed-loop configuration that incorporates the clock in a feedback loop provides more reliable synchronization. Figure 11.4-3 gives the diagram and explanatory waveforms for a representative closed-loop bit synchronizer. Here a zero-crossing detector generates a rectangular pulse with half-bit duration $T_b/2$ starting at each zero-crossing in $y(t)$. The pulsed waveform $z(t)$ then multiplies the square-wave clock signal $c(t)$ coming back from the *voltage-controlled clock* (VCC). The control voltage $v(t)$ is obtained by integrating and lowpass-filtering the product $z(t)c(t)$. The loop reaches steady state conditions when the edges of $c(t)$ and $z(t)$ are synchronized and offset by $T_b/4$, so the product has zero area and $v(t)$ remains constant. Practical implementations of this system usually feature digital components in place of the analog multiplier and integrator.

Both of the foregoing techniques work best when the zero-crossings of $y(t)$ are spaced by integer multiples of T_b. Otherwise, the synchronization will suffer from timing jitter. An additional problem arises if the message includes a long string of 1s or 0s, so $y(t)$ has no zero-crossings, and synchronism may be lost. Message scramblers discussed shortly help alleviate this problem.

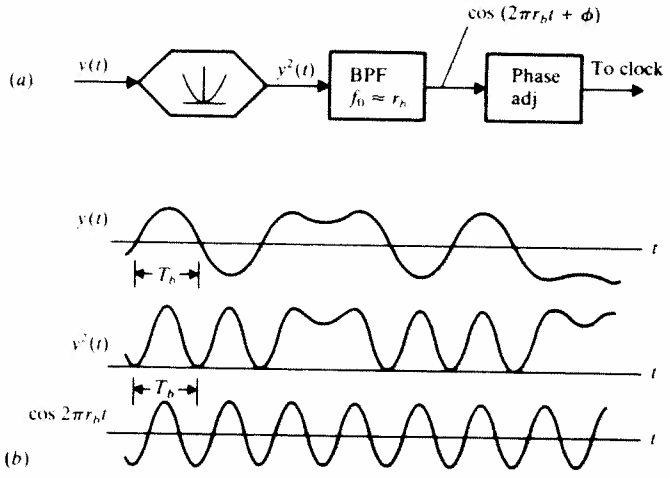


Figure 11.4-2 Bit synchronization by polar to unipolar conversion. (*a*) Block diagram; (*b*) waveforms.

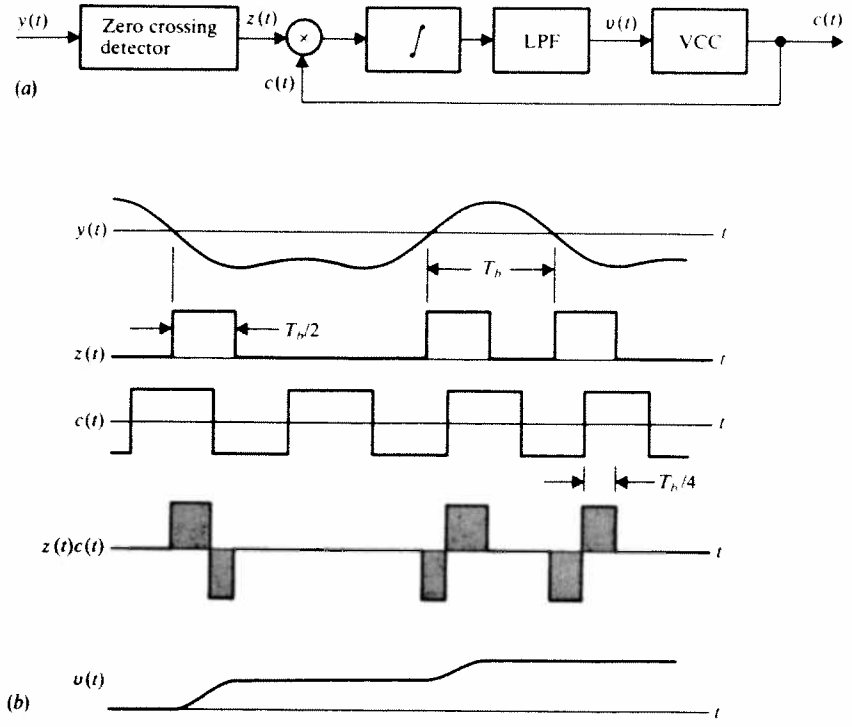


Figure 11.4-3 Closed-loop bit synchronization with a voltage-controlled clock. (*a*) Block diagram; (*b*) waveforms.

A different approach to synchronization, independent of zero-crossings, relies on the fact that a properly filtered digital signal has peaks at the optimum sampling times and is reasonably symmetric on either side. Thus, if t_k is synchronized and $\delta < T_b/2$, then

$$|y(t_k - \delta)| \approx |y(t_k + \delta)| < |y(t_k)|$$

However, a late sync signal produces the situation shown in Fig. 11.4-4*a* where $|y(t_k - \delta)| > |y(t_k + \delta)|$, while an early sync signal would result in $|y(t_k - \delta)| < |y(t_k + \delta)|$. The *early-late synchronizer* in Fig. 11.4-4*b* uses these properties to develop the control voltage for a VCC in a feedback loop. A late sync signal results in $v(t) = |y(t_k - \delta)| - |y(t_k + \delta)| > 0$, which speeds up the clock, and conversely for an early sync signal.

Scramblers and PN Sequence Generators

Scrambling is a coding operation applied to the message at the transmitter that "randomizes" the bit stream, eliminating long strings of like bits that might impair receiver synchronization. Scrambling also eliminates most periodic bit

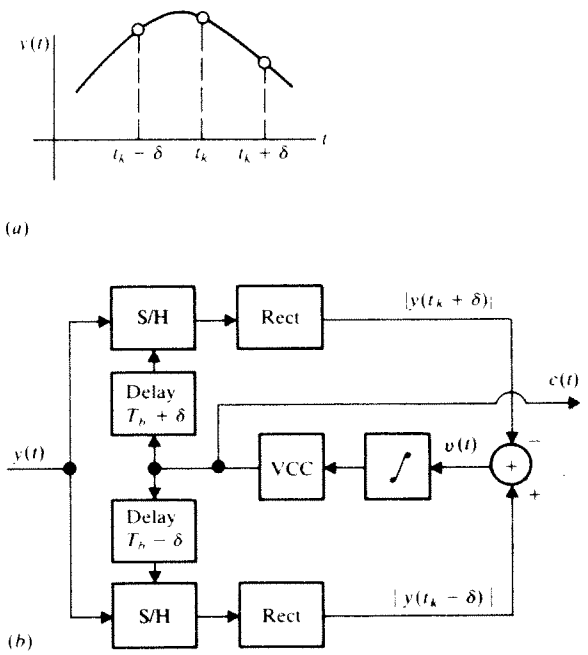


Figure 11.4-4 Early-late bit synchronization. (*a*) Waveform; (*b*) block diagram.

patterns that could produce undesirable discrete-frequency components in the power spectrum. Needless to say, the scrambled sequence must be unscrambled at the receiver so as to preserve overall *bit sequence transparency*.

Simple but effective scramblers and unscramblers are built from tapped *shift registers* having the generic form of Fig. 11.4-5, the digital counterpart of a tapped delay line. Successive bits from the binary input sequence b_k enter the register and shift from one stage to the next at each tick of the clock. The output b'_k is formed by combining the bits in the register through a set of tap gains and mod-2 adders, yielding

$$b'_k = \alpha_1 b_{k-1} \oplus \alpha_2 b_{k-2} \oplus \cdots \oplus \alpha_n b_{k-n} \tag{1}$$

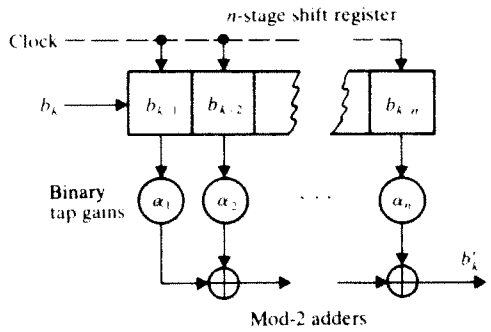


Figure 11.4-5 Tapped shift register.

The tap gains themselves are binary digits, so $\alpha_1 = 1$ simply means a direct connection while $\alpha_1 = 0$ means no connection. The symbol $\oplus$ stands for *modulo-2 addition*, defined by the properties

$$b_1 \oplus b_2 = \begin{cases} 0 & b_1 = b_2 \\ 1 & b_1 \neq b_2 \end{cases} \tag{2a}$$

and

$$b_1 \oplus b_2 \oplus b_3 = (b_1 \oplus b_2) \oplus b_3 = b_1 \oplus (b_2 \oplus b_3) \tag{2b}$$

where b_1, b_2, and b_3 are arbitrary binary digits. Mod-2 addition is implemented with exclusive-OR gates, and obeys the rules of ordinary addition except that $1 \oplus 1 = 0$.

Figure 11.4-6 shows an illustrative scrambler and unscrambler, each employing a 4-stage shift register with tap gains $\alpha_1 = \alpha_2 = 0$ and $\alpha_3 = \alpha_4 = 1$. (For clarity, we omit the clock line here and henceforth.) The binary message sequence m_k at the input to the scrambler is mod-2 added to the register output m''_k to form the scrambled message m'_k which is also fed back to the register input. Thus, $m''_k = m'_{k-3} \oplus m'_{k-4}$ and

$$m'_k = m_k \oplus m''_k \tag{3a}$$

The unscrambler has essentially the reverse structure of the scrambler and reproduces the original message sequence, since

$$\begin{aligned} m'_k \oplus m''_k &= (m_k \oplus m''_k) \oplus m''_k \\ &= m_k \oplus (m''_k \oplus m''_k) = m_k \oplus 0 = m_k \end{aligned} \tag{3b}$$

Equations (3*a*) and (3*b*) hold for any shift-register configuration as long as the scrambler and unscrambler have identical registers.

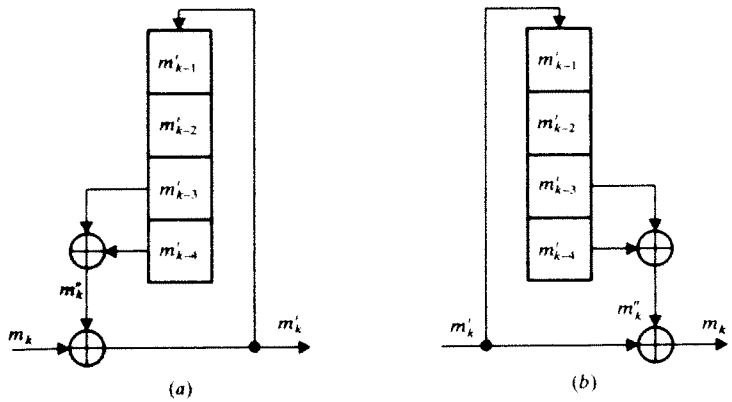


Figure 11.4-6 (*a*) Binary scrambler; (*b*) unscrambler.

Table 11.4-1

Registers contents	m'_{k-1}	0	1	0	1	0	1	1	1	1	0	0	0	1	0
	m'_{k-2}	0	0	1	0	1	0	1	1	1	1	0	0	0	1
	m'_{k-3}	0	0	0	1	0	1	0	1	1	1	1	0	0	0
	m'_{k-4}	0	0	0	0	1	0	1	0	1	1	1	1	0	0
Register output	m''_k	0	0	0	1	1	1	1	1	0	0	0	1	0	0
Input sequence	m_k	1	0	1	1	0	0	0	0	0	0	0	0	0	1
Output sequence	m'_k	1	0	1	0	1	1	1	1	0	0	0	1	0	1

The scrambling action does, of course, depend on the shift-register configuration. Table 11.4-1 portrays the scrambling produced by our illustrative scrambler when the initial state of the register is all 0s. Note that the string of nine 0s in m_k has been eliminated in m'_k. Nonetheless, there may be some specific message sequence that will result in a long string of like bits in m'_k. Of more serious concern is *error propagation* at the unscrambler, since one erroneous bit in m'_k will cause several output bit errors. Error propagation stops when the unscrambler register is full of correct bits.

Next, in preparation for the subsequent discussion of frame synchronization, we consider shift-register *sequence generation*. When a shift register has a nonzero initial state and the output is fed back to the input, the unit acts as a periodic sequence generator. Take Fig. 11.4-7, for example, where a 3-stage register with initial state 111 produces the 7-bit sequence 1110010 which repeats periodically thereafter. In general, the longest possible sequence period from a register with n stages is

$$N = 2^n - 1 \tag{4}$$

and the corresponding output is called a *maximal-length* or *pseudonoise* (PN) sequence. Figure 11.4-7 is, in fact, a PN sequence generator with $n = 3$ and $N = 7$.

The name pseudonoise comes from the *correlation* properties of PN sequences. To develop this point, let a PN sequence s_k be used to form the binary polar NRZ signal

$$s(t) = \sum_k (2s_k - 1)p(t - kT_b) \tag{5}$$

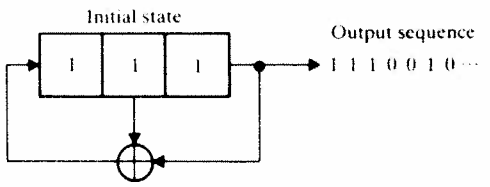


Figure 11.4-7 Shift-register sequence generator.

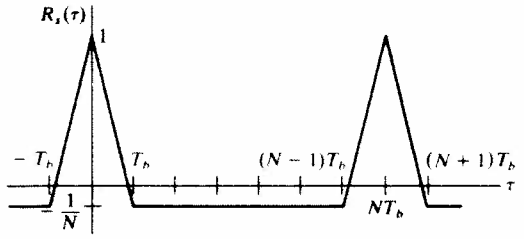


Figure 11.4-8 Autocorrelation of a PN sequence.

where $p(t)$ is a rectangular pulse and the amplitude of the kth pulse is

$$2s_k - 1 = \begin{cases} +1 & s_k = 1 \\ -1 & s_k = 0 \end{cases} \tag{6}$$

The signal $s(t)$ is deterministic and periodic, with period NT_b, and has a periodic autocorrelation function given by

$$R_s(\tau) = [(N+1)\Lambda(\tau/T_b) - 1]/N \qquad |\tau| \le NT_b/2 \tag{7}$$

which is plotted in Fig. 11.4-8. If N is very large and T_b very small, then

$$R_s(\tau) \approx T_b\,\delta(\tau) - 1/N \qquad |\tau| \le NT_b/2$$

so the PN signal acts essentially like *white noise* with a small DC component. This noise-like correlation property leads to practical applications in test instruments and radar ranging, as well as spread spectrum communication described in Sect. 9.3, and digital framing, to which we now turn.

Frame Synchronization

A digital receiver needs to know when a signal is present. Otherwise, the input noise alone may produce random output bits that could be mistaken for a message. Therefore, identifying the start of a message is one aspect of frame synchronization. Another aspect is identifying subdivisions or frames within the message. To facilitate frame synchronization, binary transmission usually includes special N-bit *sync words* as represented in Fig. 11.4-9. The initial *prefix* consists of several repetitions of the sync word, which marks the begining of transmission and allows time for bit-sync acquisition. The prefix is followed by a different codeword labeling the start of the message itself. Frames are labeled by sync words inserted periodically in the bit stream.

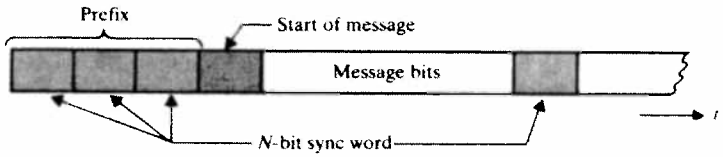


Figure 11.4-9

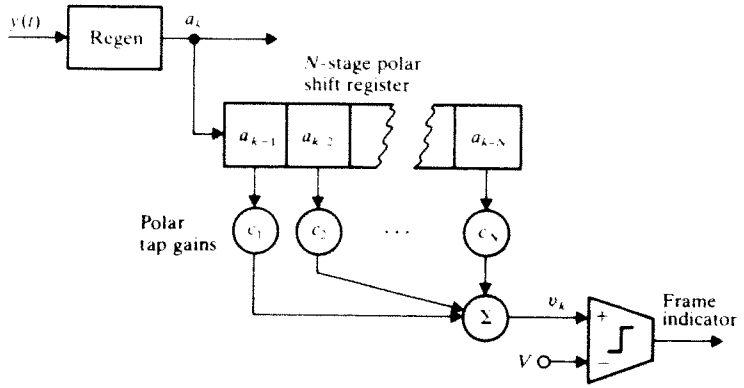


Figure 11.4-10 Frame synchronizer.

The elementary frame synchronizer in Fig. 11.4-10 is designed to detect a sync word $s_1 s_2 \cdots s_N$ whenever it appears in the regenerated sequence m_k. Output bits with the polar format

$$a_k = 2m_k - 1 = \pm 1$$

are loaded into an N-stage polar shift register having polar tap gains given by

$$c_i = 2s_{N+1-i} - 1 \tag{8}$$

This awkward-looking expression simply states that the gains equal the sync-word bits in polar form and reverse order, that is, $c_1 = 2s_N - 1$ while $c_N = 2s_1 - 1$. The tap-gain outputs are summed *algebraically* to form

$$v_k = \sum_{i=1}^{N} c_i a_{k-i} \tag{9}$$

This voltage is compared with a threshold voltage V, and the frame-sync indicator goes HIGH when $v_k > V$.

If the register word is identical to the sync word, then $a_{k-i} = c_i$ so $c_i a_{k-i} = c_i^2 = 1$ and $v_k = N$. If the register word differs from the sync word in just one bit, then $v_k = N - 2$ (why?). Setting the threshold voltage V slightly below $N - 2$ thus allows detection of error-free sync words and sync words with one bit error. Sync words with two or more errors go undetected, but that should be an unlikely event with any reasonable value of P_e. *False* frame indication occurs when N or $N - 1$ successive message bits match the sync-word bits. The probability of this event is given by the binomial model as

$$P_{ff} = \binom{N}{0}\left(\frac{1}{2}\right)^N + \binom{N}{1}\left(\frac{1}{2}\right)^{N-1}\left(\frac{1}{2}\right) = (1+N)2^{-N} \tag{10}$$

assuming equally likely 1s and 0s in the bit stream.

Further examination of Eqs. (8) and (9) reveals that the frame synchronizer calculates the *cross-correlation* between the bit stream passing through the register and the sync word, represented by the tap gains. The correlation properties of a PN sequence therefore makes it an ideal choice for the sync word. In particular, suppose the prefix consists of several periods of a PN sequence. As the prefix passes through the frame-sync register, the values of v_k will trace out the shape of $R_x(\tau)$ in Fig. 11.4-8 with peaks $v_k = N$ occurring each time the initial bit s_1 reaches the end of the register. An added advantage is the ease of PN sequence generation at the transmitter, even with large values of N. For instance, getting $P_{ff} < 10^{-6}$ in Eq. (10) requires $N > 25$, which can be accomplished with a 5-stage PN generator.

11.5 PROBLEMS

11.1-1 Sketch $x(t)$ and construct the corresponding eye pattern (without transmission distortion) for binary PAM with the data sequence 1011100010 when the signal has a unipolar format and pulse shape $p(t) = \cos^2(\pi t/2T_b)\,\Pi(t/2T_b)$.

11.1-2 Do Prob. 11.1-1 with $p(t) = \cos(\pi t/2T_b)\,\Pi(t/2T_b)$.

11.1-3 Do Prob. 11.1-1 with a polar format and $p(t) = \cos(\pi t/2T_b)\,\Pi(t/2T_b)$.

11.1-4 Do Prob. 11.1-1 with a bipolar format and $p(t) = \Lambda(t/T_b)$.

11.1-5 Do Prob. 11.1-1 with a bipolar format.

11.1-6 Modify Table 11.1-1 for an *octal* signal ($M = 8$).

11.1-7 A certain computer generates binary words, each consisting of 16 bits, at the rate 20,000 words per second. (*a*) Find the bandwidth required to transmit the output as a binary PAM signal. (*b*) Find M so that the output could be transmitted as an M-ary signal on a channel having $B = 60$ kHz.

11.1-8 A certain digital tape reader produces 3000 symbols per second, and there are 128 different symbols. (*a*) Find the bandwidth required to transmit the output as a binary PAM signal. (*b*) Find M so that the output could be transmitted as an M-ary signal on a telephone link having $B = 3$ kHz.

11.1-9 Suppose a digital signal with *pulse-position modulation* (PPM) is constructed by dividing the symbol interval into $M + 1$ equal time slots, allowing for M nonoverlapping pulse positions and a one-slot guard time. Develop an expression for the minimum bandwidth B needed to transmit M-ary data at a fixed bit rate r_b. Then tabulate B/r_b for a few values of $M = 2^n$ and determine the most bandwidth-efficient choice of M.

11.1-10 Binary data is transmitted as a unipolar signal with $A = 1$ and $p(t) = u(t + T_b) - u(t)$. The transmission system's step response is $g(t) = K_0(1 - e^{-bt})u(t)$, where $b = 2/T_b$. (*a*) Sketch $\tilde{p}(t)$ and find K_0 such that $\tilde{p}(0) = 1$. (*b*) Sketch $y(t)$ for the data sequence 10110 and evaluate $y(t)$ and the ISI at the optimum sampling times.

11.1-11 Do Prob. 11.1-10 for a polar signal with $A/2 = 1$ and a transmission system having $b = 1/T_b$.

11.1-12 Consider digital transmission with a gaussian pulse shape $p(t) = \exp[-\pi(bt)^2]$, which is neither timelimited nor bandlimited and does not have periodic zero crossings. Let $p(kD) \le 0.01$ for $k \neq 0$, to limit the ISI, and let the bandwidth B be such that $P(f) \le 0.01P(0)$ for $|f| > B$. Find the resulting relationship between r and B.

11.1-13 Find and sketch the power spectrum of a binary PAM signal with polar RZ format and rectangular pulses, assuming independent and equiprobable message bits. Then show that the time-domain and frequency-domain calculations of $\overline{x^2}$ are in agreement.