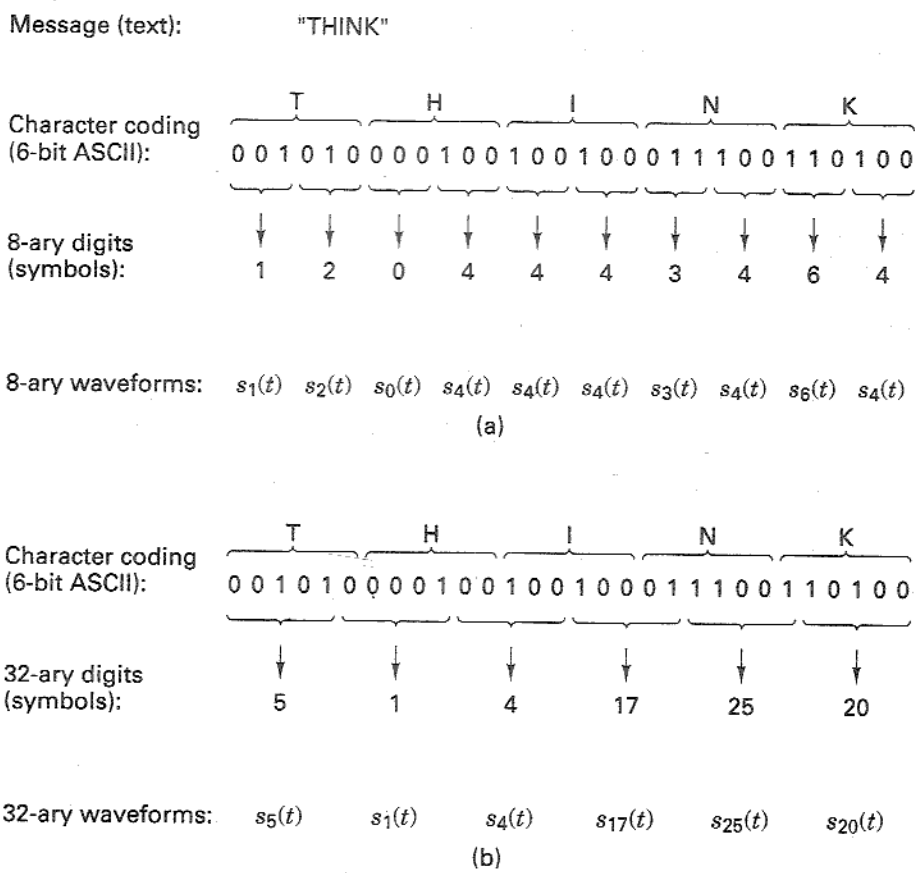

Figure 2.5 Messages, characters, and symbols. (a) 8-ary example. (b) 32-ary example.

represents the remaining $\frac{1}{6}$ of the character "T" and $\frac{4}{6}$ of the next character, "H," and so on. It is not necessary that the characters be partitioned more aesthetically. The system sees the characters as a string of digits to be transmitted; only the end user (or the user's teleprinter machine) ascribes textual meaning to the final delivered sequence of bits. In this 32-ary case, a transmitter needs a repertoire of 32 waveforms $s_i(t)$, where $i = 1, \ldots, 32$, one for each possible symbol that may be transmitted. The final row of the figure lists the six waveforms that a 32-ary modulating system transmits to represent the textual message "THINK."

2.4 FORMATTING ANALOG INFORMATION

If the information is analog, it cannot be character encoded as in the case of textual data; the information must first be transformed into a digital format. The process of transforming an analog waveform into a form that is compatible with a digital com-

munication system starts with sampling the waveform to produce a discrete pulse-amplitude-modulated waveform, as described below.

2.4.1 The Sampling Theorem

The link between an analog waveform and its sampled version is provided by what is known as the *sampling process.* This process can be implemented in several ways, the most popular being the *sample-and-hold* operation. In this operation, a switch and storage mechanism (such as a transistor and a capacitor, or a shutter and a filmstrip) form a sequence of samples of the continuous input waveform. The output of the sampling process is called *pulse amplitude modulation* (PAM) because the successive output intervals can be described as a sequence of pulses with amplitudes derived from the input waveform samples. The analog waveform can be approximately retrieved from a PAM waveform by simple low-pass filtering. An important question: how closely can a filtered PAM waveform approximate the original input waveform? This question can be answered by reviewing the *sampling theorem,* which states the following [1]: A bandlimited signal having no spectral components above f_m hertz can be determined uniquely by values sampled at uniform intervals of

$$T_s \leq \frac{1}{2f_m} \text{ sec} \tag{2.1}$$

This particular statement is also known as the *uniform sampling theorem.* Stated another way, the upper limit on T_s can be expressed in terms of the sampling rate, denoted $f_s = 1/T_s$. The restriction, stated in terms of the sampling rate, is known as the *Nyquist criterion.* The statement is

$$f_s \geq 2f_m \tag{2.2}$$

The sampling rate $f_s = 2f_m$ is also called the *Nyquist rate.* The Nyquist criterion is a theoretically sufficient condition to allow an analog signal to be *reconstructed completely* from a set of uniformly spaced discrete-time samples. In the sections that follow, the validity of the sampling theorem is demonstrated using different sampling approaches.

2.4.1.1 Impulse Sampling

Here we demonstrate the validity of the sampling theorem using the frequency convolution property of the Fourier transform. Let us first examine the case of *ideal sampling* with a sequence of unit impulse functions. Assume an analog waveform, $x(t)$, as shown in Figure 2.6a, with a Fourier transform, $X(f)$, which is zero outside the interval $(-f_m < f < f_m)$, as shown in Figure 2.6b. The sampling of $x(t)$ can be viewed as the product of $x(t)$ with a periodic train of unit impulse functions $x_\delta(t)$, shown in Figure 2.6c and defined as

$$x_\delta(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT_s) \tag{2.3}$$

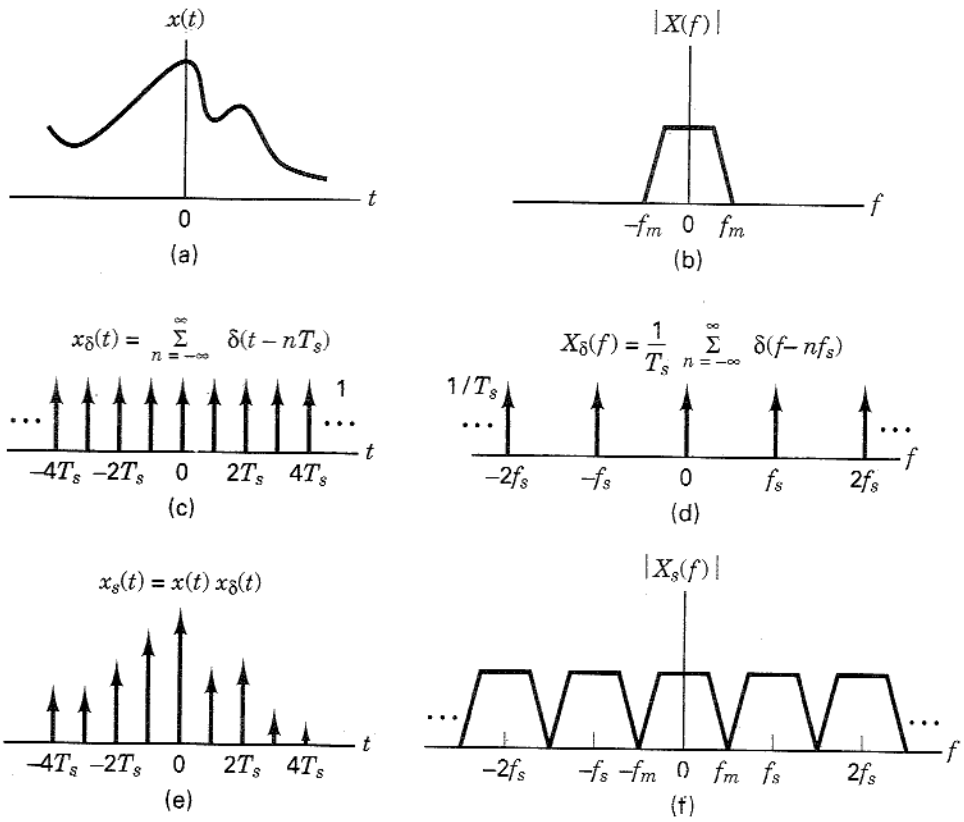

Figure 2.6 Sampling theorem using the frequency convolution property of the Fourier transform.

where T_s is the sampling period and $\delta(t)$ is the unit impulse or Dirac delta function defined in Section 1.2.5. Let us choose $T_s = 1/2f_m$, so that the Nyquist criterion is just satisfied.

The *sifting property* of the impulse function (see Section A.4.1) states that

$$x(t)\delta(t - t_0) = x(t_0)\delta(t - t_0) \tag{2.4}$$

Using this property, we can see that $x_s(t)$, the sampled version of $x(t)$ shown in Figure 2.6e, is given by

$$\begin{aligned} x_s(t) = x(t)x_\delta(t) &= \sum_{n=-\infty}^{\infty} x(t)\delta(t - nT_s) \\ &= \sum_{n=-\infty}^{\infty} x(nT_s)\delta(t - nT_s) \end{aligned} \tag{2.5}$$

Using the *frequency convolution property* of the Fourier transform (see Section A.5.3), we can transform the time-domain product $x(t)x_\delta(t)$ of Equation (2.5) to the frequency-domain convolution $X(f) * X_\delta(f)$, where

$$X_\delta(f) = \frac{1}{T_s} \sum_{n=-\infty}^{\infty} \delta(f - nf_s) \tag{2.6}$$

is the Fourier transform of the impulse train $x_\delta(t)$ and where $f_s = 1/T_s$ is the sampling frequency. Notice that the Fourier transform of an impulse train is another impulse train; the values of the periods of the two trains are reciprocally related to one another. Figures 2.6c and d illustrate the impulse train $x_\delta(t)$ and its Fourier transform $X_\delta(f)$, respectively.

Convolution with an impulse function simply shifts the original function as follows:

$$X(f) * \delta(f - nf_s) = X(f - nf_s) \tag{2.7}$$

We can now solve for the transform $X_s(f)$ of the sampled waveform:

$$\begin{aligned} X_s(f) = X(f) * X_\delta(f) &= X(f) * \left[\frac{1}{T_s} \sum_{n=-\infty}^{\infty} \delta(f - nf_s) \right] \\ &= \frac{1}{T_s} \sum_{n=-\infty}^{\infty} X(f - nf_s) \end{aligned} \tag{2.8}$$

We therefore conclude that within the original bandwidth, the spectrum $X_s(f)$ of the sampled signal $x_s(t)$ is, to within a constant factor $(1/T_s)$, exactly the same as that of $x(t)$. In addition, the spectrum repeats itself periodically in frequency every f_s hertz. The sifting property of an impulse function makes the convolving of an impulse train with another function easy to visualize. The impulses act as sampling functions. Hence, convolution can be performed graphically by sweeping the impulse train $X_\delta(f)$ in Figure 2.6d past the transform $|X(f)|$ in Figure 2.6b. This sampling of $|X(f)|$ at each step in the sweep replicates $|X(f)|$ at each of the frequency positions of the impulse train, resulting in $|X_s(f)|$, shown in Figure 2.6f.

When the sampling rate is chosen, as it has been here, such that $f_s = 2f_m$, each spectral replicate is separated from each of its neighbors by a frequency band exactly equal to f_s hertz, and the analog waveform can theoretically be completely recovered from the samples, by the use of filtering. However, a filter with infinitely steep sides would be required. It should be clear that if $f_s > 2f_m$, the replications will move farther apart in frequency, as shown in Figure 2.7a, making it easier to perform the filtering operation. A typical low-pass filter characteristic that might be used to separate the baseband spectrum from those at higher frequencies is shown in the figure. When the sampling rate is reduced, such that $f_s < 2f_m$, the replications will overlap, as shown in Figure 2.7b, and some information will be lost. The phenomenon, the result of undersampling (sampling at too low a rate), is called *aliasing*. The Nyquist rate, $f_s = 2f_m$, is the sampling rate below which aliasing occurs; to avoid aliasing, the Nyquist criterion, $f_s \geq 2f_m$, must be satisfied.

As a matter of practical consideration, neither waveforms of engineering interest nor realizable bandlimiting filters are strictly bandlimited. Perfectly bandlimited signals do not occur in nature (see Section 1.7.2); thus, realizable signals, even though we may think of them as bandlimited, always contain some aliasing. These signals and filters can, however, be considered to be "essentially" bandlimited. By

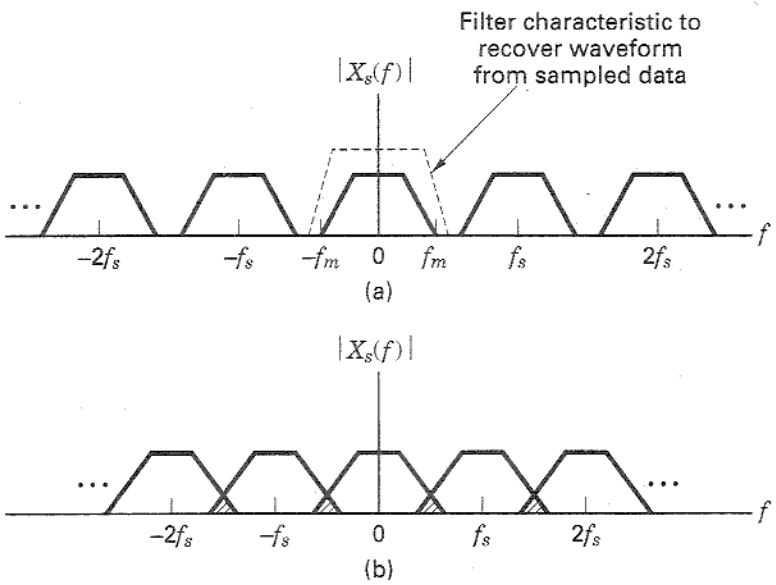

Figure 2.7 Spectra for various sampling rates. (a) Sampled spectrum ($f_s > 2f_m$). (b) Sampled spectrum ($f_s < 2f_m$).

this we mean that a bandwidth can be determined beyond which the spectral components are attenuated to a level that is considered negligible.

2.4.1.2 Natural Sampling

Here we demonstrate the validity of the sampling theorem using the frequency shifting property of the Fourier transform. Although instantaneous sampling is a convenient model, a more practical way of accomplishing the sampling of a bandlimited analog signal $x(t)$ is to multiply $x(t)$, shown in Figure 2.8a, by the pulse train or switching waveform $x_p(t)$, shown in Figure 2.8c. Each pulse in $x_p(t)$ has width T and amplitude $1/T$. Multiplication by $x_p(t)$ can be viewed as the opening and closing of a switch. As before, the sampling frequency is designated f_s, and its reciprocal, the time period between samples, is designated T_s. The resulting sampled-data sequence, $x_s(t)$, is illustrated in Figure 2.8e and is expressed as

$$x_s(t) = x(t)x_p(t) \tag{2.9}$$

The sampling here is termed *natural sampling,* since the top of each pulse in the $x_s(t)$ sequence retains the shape of its corresponding analog segment during the pulse interval. Using Equation (A.13), we can express the periodic pulse train as a Fourier series in the form

$$x_p(t) = \sum_{n=-\infty}^{\infty} c_n e^{j\,2\pi n f_s t} \tag{2.10}$$

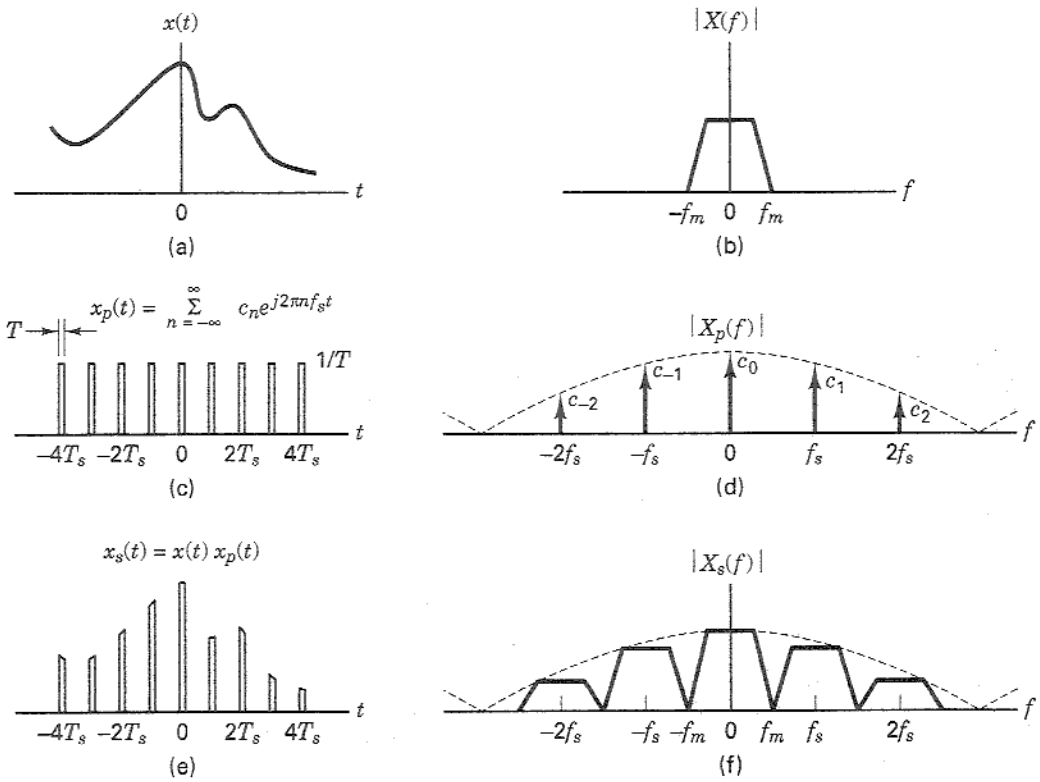

Figure 2.8 Sampling theorem using the frequency shifting property of the Fourier transform.

where the sampling rate, $f_s = 1/T_s$, is chosen equal to $2f_m$, so that the Nyquist criterion is just satisfied. From Equation (A.24), $c_n = (1/T_s)$ sinc (nT/T_s), where T is the pulse width, $1/T$ is the pulse amplitude, and

$$\text{sinc } y = \frac{\sin \pi y}{\pi y}$$

The envelope of the magnitude spectrum of the pulse train, seen as a dashed line in Figure 2.8d, has the characteristic sinc shape. Combining Equations (2.9) and (2.10) yields

$$x_s(t) = x(t) \sum_{n=-\infty}^{\infty} c_n e^{j\,2\pi n f_s t} \tag{2.11}$$

The transform $X_s(f)$ of the sampled waveform is found as follows:

$$X_s(f) = \mathcal{F}\left\{x(t) \sum_{n=-\infty}^{\infty} c_n e^{j\,2\pi n f_s t}\right\} \tag{2.12}$$

For linear systems, we can interchange the operations of summation and Fourier transformation. Therefore, we can write

$$X_s(f) = \sum_{n=-\infty}^{\infty} c_n \mathcal{F}\{x(t)e^{j\,2\pi n f_s t}\} \tag{2.13}$$

Using the *frequency translation* property of the Fourier transform (see Section A.3.2), we solve for $X_s(f)$ as follows:

$$X_s(f) = \sum_{n=-\infty}^{\infty} c_n X(f - nf_s) \tag{2.14}$$

Similar to the unit impulse sampling case, Equation (2.14) and Figure 2.8f illustrate that $X_s(f)$ is a replication of $X(f)$, periodically repeated in frequency every f_s hertz. In this natural-sampled case, however, we see that $X_s(f)$ is weighted by the Fourier series coefficients of the pulse train, compared with a constant value in the impulse-sampled case. It is satisfying to note that *in the limit,* as the pulse width, T, approaches zero, c_n approaches $1/T_s$ for all n (see the example that follows), and Equation (2.14) converges to Equation (2.8).

Example 2.1 Comparison of Impulse Sampling and Natural Sampling

Consider a given waveform $x(t)$ with Fourier transform $X(f)$. Let $X_{s1}(f)$ be the spectrum of $x_{s1}(t)$, which is the result of sampling $x(t)$ with a unit impulse train $x_\delta(t)$. Let $X_{s2}(f)$ be the spectrum of $x_{s2}(t)$, the result of sampling $x(t)$ with a pulse train $x_p(t)$ with pulse width T, amplitude $1/T$, and period T_s. Show that in the limit, as T approaches zero, $X_{s1}(f) = X_{s2}(f)$.

Solution

From Equation (2.8),

$$X_{s1}(f) = \frac{1}{T_s} \sum_{n=-\infty}^{\infty} X(f - nf_s)$$

and from Equation (2.14),

$$X_{s2}(f) = \sum_{n=-\infty}^{\infty} c_n X(f - nf_s)$$

As the pulse with $T \to 0$, and the pulse amplitude approaches infinity (the area of the pulse remains unity), $x_p(t) \to x_\delta(t)$. Using Equation (A.14), we can solve for c_n in the limit as follows:

$$\begin{aligned} c_n &= \lim_{T \to 0} \frac{1}{T_s} \int_{-T_s/2}^{T_s/2} x_p(t) e^{-j\,2\pi n f_s t}\, dt \\ &= \frac{1}{T_s} \int_{-T_s/2}^{T_s/2} x_\delta(t) e^{-j\,2\pi n f_s t}\, dt \end{aligned}$$

Since, within the range of integration, $-T_s/2$ to $T_s/2$, the only contribution of $x_\delta(t)$ is that due to the impulse at the origin, we can write

$$c_n = \frac{1}{T_s}\int_{-T_s/2}^{T_s/2} \delta(t)e^{-j\,2\pi n f_s t}\,dt = \frac{1}{T_s}$$

Therefore, in the limit, $X_{s1}(f) = X_{s2}(f)$ for all n.

2.4.1.3 Sample-and-Hold Operation

The simplest and thus most popular sampling method, *sample and hold,* can be described by the convolution of the sampled pulse train, $[x(t)x_\delta(t)]$, shown in Figure 2.6e, with a unity amplitude rectangular pulse $p(t)$ of pulse width T_s. This time, convolution results in the *flattop* sampled sequence

$$\begin{aligned} x_s(t) &= p(t) * [x(t)x_\delta(t)] \\ &= p(t) * \left[x(t)\sum_{n=-\infty}^{\infty}\delta(t - nT_s)\right] \end{aligned} \tag{2.15}$$

The Fourier transform, $X_s(f)$, of the time convolution in Equation (2.15) is the frequency-domain product of the transform $P(f)$ of the rectangular pulse and the periodic spectrum, shown in Figure 2.6f, of the impulse-sampled data:

$$\begin{aligned} X_s(f) &= P(f)\mathcal{F}\left\{x(t)\sum_{n=-\infty}^{\infty}\delta(t - nT_s)\right\} \\ &= P(f)\left\{X(f) * \left[\frac{1}{T_s}\sum_{n=-\infty}^{\infty}\delta(f - nf_s)\right]\right\} \\ &= P(f)\frac{1}{T_s}\sum_{n=-\infty}^{\infty}X(f - nf_s) \end{aligned} \tag{2.16}$$

Here, $P(f)$ is of the form T_s sinc fT_s. The effect of this product operation results in a spectrum similar in appearance to the natural-sampled example presented in Figure 2.8f. The most obvious effect of the hold operation is the significant attenuation of the higher-frequency spectral replicates (compare Figure 2.8f to Figure 2.6f), which is a desired effect. Additional analog postfiltering is usually required to finish the filtering process by further attenuating the residual spectral components located at the multiples of the sample rate. A secondary effect of the hold operation is the nonuniform spectral gain $P(f)$ applied to the desired baseband spectrum shown in Equation (2.16). The postfiltering operation can compensate for this attenuation by incorporating the inverse of $P(f)$ over the signal passband.

2.4.2 Aliasing

Figure 2.9 is a detailed view of the positive half of the baseband spectrum and one of the replicates from Figure 2.7b. It illustrates aliasing in the frequency domain. The overlapped region, shown in Figure 2.9b, contains that part of the spectrum which is aliased due to *undersampling.* The aliased spectral components represent ambiguous data that appear in the frequency band between $(f_s - f_m)$ and f_m. Figure 2.10 illustrates that a higher sampling rate f'_s, can eliminate the aliasing by separat-

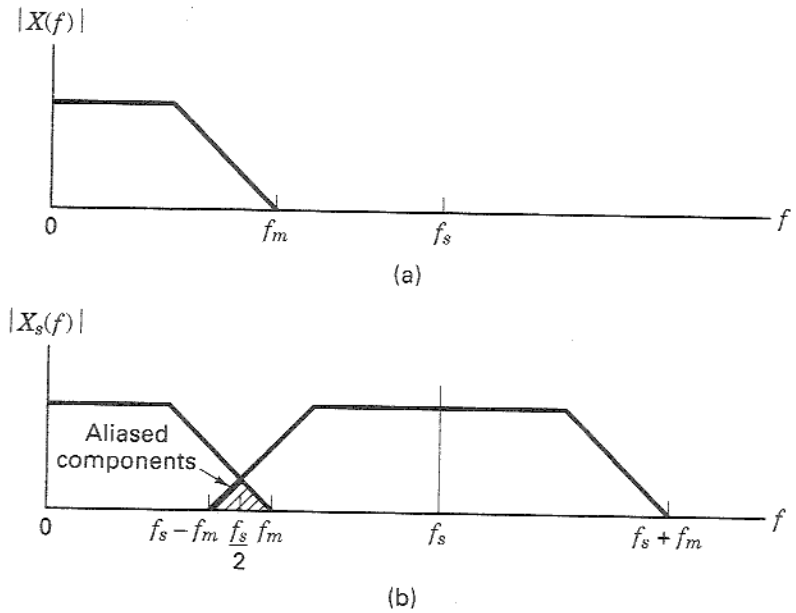

Figure 2.9 Aliasing in the frequency domain. (a) Continuous signal spectrum. (b) Sampled signal spectrum.

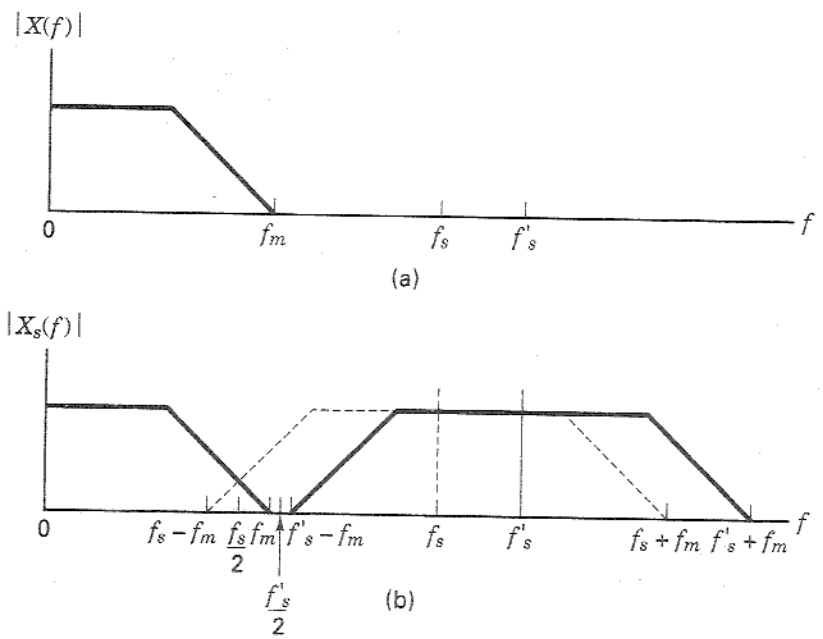

Figure 2.10 Higher sampling rate eliminates aliasing. (a) Continuous signal spectrum. (b) Sampled signal spectrum.

ing the spectral replicates; the resulting spectrum in Figure 2.10b corresponds to the case in Figure 2.7a. Figures 2.11 and 2.12 illustrate two ways of eliminating aliasing using *antialiasing filters.* In Figure 2.11 the analog signal is *prefiltered* so that the new maximum frequency, f'_m, is reduced to $f_s/2$ or less. Thus there are no aliased components seen in Figure 2.11b, since $f_s > 2f'_m$. Eliminating the aliasing terms prior to sampling is good engineering practice. When the signal structure is well known, the aliased terms can be eliminated after sampling, with a low-pass filter operating on the sampled data [2]. In Figure 2.12 the aliased components are removed by *postfiltering* after sampling; the filter cutoff frequency, f''_m, removes the aliased components; f''_m needs to be less than $(f_s - f_m)$. Notice that the filtering techniques for eliminating the aliased portion of the spectrum in Figures 2.11 and 2.12 *will result in a loss* of some of the signal information. For this reason, the sample rate, cutoff bandwidth, and filter type selected for a particular signal bandwidth are all interrelated.

Realizable filters require a nonzero bandwidth for the transition between the passband and the required out-of-band attenuation. This is called the *transition bandwidth.* To minimize the system sample rate, we desire that the antialiasing filter have a small transition bandwidth. Filter complexity and cost rise sharply with narrower transition bandwidth, so a trade-off is required between the cost of a small transition bandwidth and the costs of the higher sampling rate, which are those of more storage and higher transmission rates. In many systems the answer has been to make the transition bandwidth between 10 and 20% of the signal band-

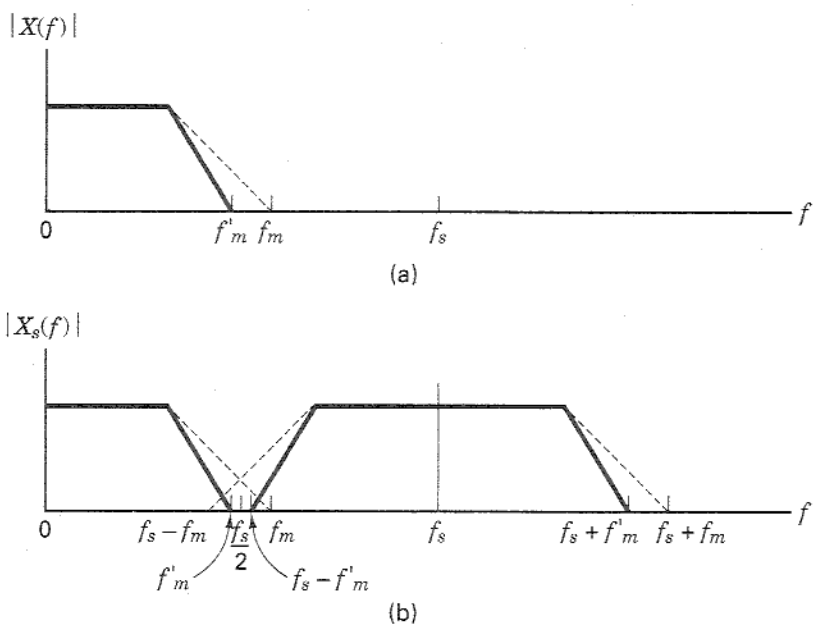

Figure 2.11 Sharper-cutoff filters eliminate aliasing. (a) Continuous signal spectrum. (b) Sampled signal spectrum.

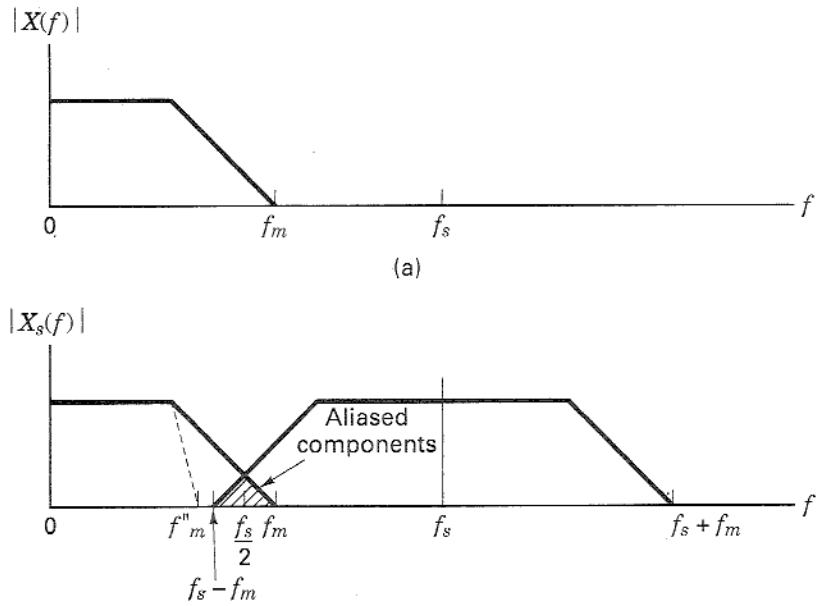

Figure 2.12 Postfilter eliminates aliased portion of spectrum. (a) Continuous signal spectrum. (b) Sampled signal spectrum.

width. If we account for the 20% transition bandwidth of the antialiasing filter, we have an *engineer's version* of the Nyquist sampling rate:

$$f_s \geq 2.2f_m \tag{2.17}$$

Figure 2.13 provides some insight into aliasing as seen in the time domain. The sampling instants of the solid-line sinusoid have been chosen so that the sinusoidal signal is undersampled. Notice that the resulting ambiguity allows one to draw a totally different (dashed-line) sinusoid, following the undersampled points.

Example 2.2 Sampling Rate for a High-Quality Music System

We wish to produce a high-quality digitization of a 20-kHz bandwidth music source. We are to determine a reasonable sample rate for this source. By the engineer's version of the Nyquist rate, in Equation (2.17), the sampling rate should be greater than 44.0 ksamples/s. As a matter of comparison, the standard sampling rate for the compact disc digital audio player is 44.1 ksamples/s, and the standard sampling rate for studio-quality audio is 48.0 ksamples/s.

2.4.3 Why Oversample?

Oversampling is the most economic solution for the task of transforming an analog signal to a digital signal, or the reverse, transforming a digital signal to an analog signal. This is so because signal processing performed with high performance ana-

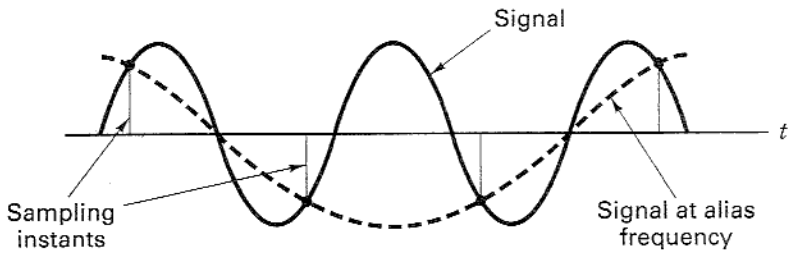

Figure 2.13 Alias frequency generated by sub-Nyquist sampling rate.

log equipment is typically much more costly than using digital signal processing equipment to perform the same task. Consider the task of transforming analog signals to digital signals. When this task is performed without the benefit of over-sampling, the process is characterized by three simple steps, performed in the order that follows.

Without Oversampling

1. The signal passes through a high performance analog lowpass filter to limit its bandwidth.
2. The filtered signal is sampled at the Nyquist rate for the (approximated) bandlimited signal. As described in Section 1.7.2, a strictly bandlimited signal is not realizable.
3. The samples are processed by an analog-to-digital converter that maps the continuous-valued samples to a finite list of discrete output levels.

When this task is performed with the benefit of over-sampling, the process is best described as five simple steps, performed in the order that follows.

With Oversampling

1. The signal is passed through a low performance (less costly) analog low-pass filter (prefilter) to limit its bandwidth.
2. The pre-filtered signal is sampled at the (now higher) Nyquist rate for the (approximated) bandlimited signal.
3. The samples are processed by an analog-to-digital converter that maps the continuous-valued samples to a finite list of discrete output levels.
4. The digital samples are then processed by a high performance digital filter to reduce the bandwidth of the digital samples.
5. The sample rate at the output of the digital filter is reduced in proportion to the bandwidth reduction obtained by this digital filter.

The next two sections examine the benefits of over-sampling.

2.4.3.1 Analog Filtering, Sampling, and Analog to Digital Conversion

The analog filter that limits the bandwidth of an input signal has a passband frequency equal to the signal bandwidth, followed by a transition to a stop band. The bandwidth of the transition region results in an increase in bandwidth of the output signal by some amount f_t. The Nyquist rate f_s for the filtered output, nominally equal to $2f_m$ (twice the highest frequency in the sampled signal) must now be increased to $2f_m + f_t$. The transition bandwidth of the filter represents an overhead in the sampling process. This additional spectral interval does not represent useful signal bandwidth but rather protects the signal bandwidth by reserving a spectral region for the aliased spectrum due to the sampling process. The aliasing stems from the fact that real signals cannot be strictly bandlimited. Typical transition bandwidths represent a 10- to 20-percent increase of the sample rate relative to that dictated by the Nyquist criterion. Examples of this overhead are seen in the compact disc (CD) digital audio system, for which the two-sided bandwidth is 40 kHz and the sample rate is 44.1 kHz, and also in the digital audio tape (DAT) system, which also has a two-sided bandwidth of 40 kHz with a sample rate of 48.0 kHz.

Our intuition and initial impulse is to keep the sample rate as low as possible by building analog filters with narrow transition bandwidths. However, analog filters can exhibit two undesirable characteristics. First, they can exhibit distortion (nonlinear phase versus frequency) due to narrow transition bandwidths. Second, the cost can be high because narrow transition bandwidths dictate high-order filters (see Section 1.6.3.2) requiring a large number of high-quality components. Our quandary is that we wish to operate the sampler at the lowest possible rate to reduce the data-storage cost. To meet this goal we might build a sophisticated analog filter with a narrow transition bandwidth. But such a filter is not only expensive, it also distorts the very signal it has been designed to protect (from undesired aliasing).

The solution (oversampling) is elegant—having been given a problem that we can't solve, we convert it to one that we can solve. We elect to use a low-cost, less sophisticated analog prefilter to limit the bandwidth of the input signal. This analog filter has been simplified by choosing a wider transition bandwidth. With a wider transition bandwidth, the required sample rate must now be increased to accommodate this larger spectrum. We typically start by selecting the higher sample rate to be 4 times the original sample rate, and then we design the analog filter to have a transition bandwidth that matches the increased sample rate. As an example, rather than sampling a CD signal at 44.1 kHz with a transition bandwidth of 4.1 kHz implemented with a sophisticated 10th order elliptic filter (implying that the filter includes 10 energy storage elements, such as capacitors and inductors), we might choose the option to employ oversampling. In that case, we could operate the sampler at 176.4 kHz with a transition bandwidth of 136.4 kHz implemented with a simpler 4th-order elliptic filter (having only 4 energy storage elements).

2.4.3.2 Digital Filtering and Resampling

Now that we have the sampled data, with its higher-than-desired sample rate, we pass the sampled data through a high-performance, low-cost, digital filter to perform the desired anti-alias filtering. The digital filter can realize the narrow

transition bandwidth without the distortion associated with analog filters, and it can operate at low cost. We next reduce the sample rate of the signal (resample) after the digital filtering operation that had reduced the transition bandwidth. Good digital signal processing techniques combine the filtering and the resampling in a single structure.

Now we address a system consideration to further improve the quality of the data collection process. The analog prefilter induces some amplitude and phase distortion. We know precisely what this distortion is, and we design the digital filter so that it not only completes the anti-aliasing task of the analog prefilter, but also compensates for its gain and phase distortion. The composite response can be made as good as we want it to be. Thus we obtain a collected signal of higher quality (less distortion) at reduced cost. Digital signal processing hardware, an extension of the computer industry, is characterized by significantly lower prices each year, which has not been the case with analog processing.

In a similar fashion, oversampling is employed in the process of converting the digital signal to an analog signal (DAC). The analog filter following the DAC suffers from distortion if it has a sharp transition bandwidth. But the transition bandwidth will not be narrow if the output data presented to the DAC has been digitally oversampled.

2.4.4 Signal Interface for a Digital System

Let us examine four ways in which analog source information can be described. Figure 2.14 illustrates the choices. Let us refer to the waveform in Figure 2.14a as the *original analog waveform.* Figure 2.14b represents a sampled version of the original waveform, typically referred to as *natural-sampled data* or PAM (pulse amplitude modulation). Do you suppose that the sampled data in Figure 2.14b are compatible with a digital system? No, they are not, because the amplitude of each natural sample still has an infinite number of possible values; a digital system deals with a finite number of values. Even if the sampling is flat-top sampling, the possible pulse values form an infinite set, since they reflect all the possible values of the continuous analog waveform. Figure 2.14c illustrates the original waveform represented by discrete pulses. Here the pulses have flat tops *and* the pulse amplitude values are limited to a finite set. Each pulse is expressed as a level from a finite number of predetermined levels; each such level can be represented by a symbol from a finite alphabet. The pulses in Figure 2.14c are referred to as *quantized samples;* such a format is the obvious choice for interfacing with a digital system. The format in Figure 2.14d may be construed as the output of a sample-and-hold circuit. When the sample values are quantized to a finite set, this format can also interface with a digital system. After quantization, the analog waveform can still be recovered, but not precisely; improved reconstruction fidelity of the analog waveform can be achieved by increasing the number of quantization levels (requiring increased system bandwidth). Signal distortion due to quantization is treated in the following sections (and later in Chapter 13).

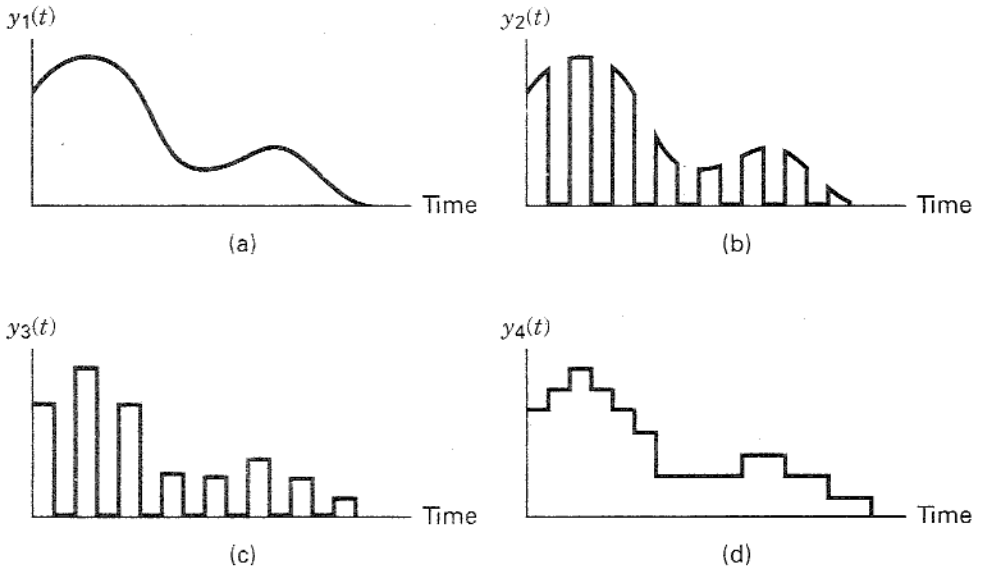

Figure 2.14 Amplitude and time coordinates of source data. (a) Original analog waveform. (b) Natural-sampled data. (c) Quantized samples. (d) Sample and hold.

2.5 SOURCES OF CORRUPTION

The analog signal recovered from the sampled, quantized, and transmitted pulses will contain corruption from several sources. The sources of corruption are related to (1) sampling and quantizing effects, and (2) channel effects. These effects are considered in the sections that follow.

2.5.1 Sampling and Quantizing Effects

2.5.1.1 Quantization Noise

The distortion inherent in quantization is a round-off or truncation error. The process of encoding the PAM signal into a quantized PAM signal involves discarding some of the original analog information. This distortion, introduced by the need to approximate the analog waveform with quantized samples, is referred to as *quantization noise;* the amount of such noise is inversely proportional to the number of levels employed in the quantization process. (The signal-to-noise ratio of quantized pulses is treated in Sections 2.5.3 and 13.2.)

2.5.1.2 Quantizer Saturation

The quantizer (or analog-to-digital converter) allocates L levels to the task of approximating the continuous range of inputs with a finite set of outputs. The range of inputs for which the difference between the input and output is small is

that there are some communication applications where delay is permissible. For example, consider the transmission of planetary images from a spacecraft. The Galileo project, launched in 1989, was on such a mission to photograph and transmit images of the planet Jupiter. The Galileo spacecraft arrived at its Jupiter destination in 1995. The journey took several years; therefore, any excess signal delay of several minutes (or hours or days) would certainly not be a problem. In such cases, the cost of more quantization levels and greater fidelity need not be bandwidth; it can be time delay.

In Figure 2.1, the term "PCM" appears in two places. First, it is a formatting topic, since the process of analog-to-digital (A/D) conversion involves sampling, quantization, and ultimately yields binary digits via the conversion of quantized PAM to PCM. Here, PCM digits are just binary numbers—a baseband carrier wave has not yet been discussed. The second appearance of PCM in Figure 2.1 is under the heading *Baseband Signaling*. Here, we list various PCM waveforms (line codes) that can be used to "carry" the PCM digits. Therefore, note that the difference between PCM and a PCM waveform is that the former represents a bit sequence, and the latter represents a particular waveform conveyance of that sequence.

2.7 UNIFORM AND NONUNIFORM QUANTIZATION

2.7.1 Statistics of Speech Amplitudes

Speech communication is a very important and specialized area of digital communications. Human speech is characterized by unique statistical properties; one such property is illustrated in Figure 2.17. The abscissa represents speech signal magnitudes, normalized to the root-mean-square (rms) value of such magnitudes through a typical communication channel, and the ordinate is probability. For most voice

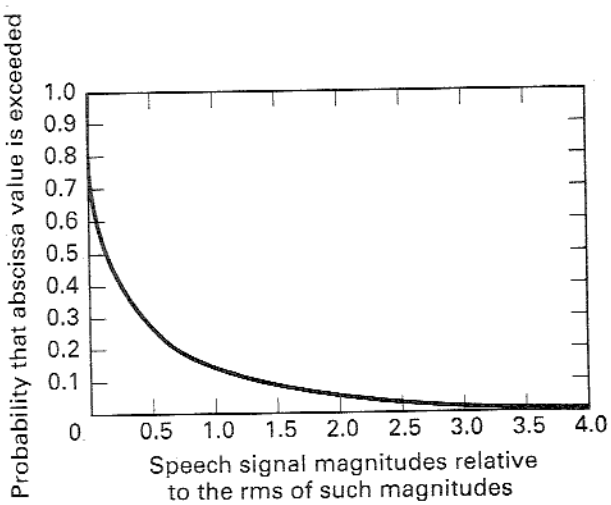

Figure 2.17 Statistical distribution of single-talker speech signal magnitudes.

communication channels, very low speech volumes predominate; 50% of the time, the voltage characterizing detected speech energy is less than one-fourth of the rms value. Large amplitude values are relatively rare; only 15% of the time does the voltage exceed the rms value. We see from Equation (2.18b) that the quantization noise depends on the step size (size of the quantile interval). When the steps are uniform in size the quantization is known as *uniform quantization.* Such a system would be wasteful for speech signals; many of the quantizing steps would rarely be used. In a system that uses equally spaced quantization levels, the quantization noise is the same for all signal magnitudes. Therefore, with uniform quantization, the signal-to-noise (SNR) is worse for low-level signals than for high-level signals. *Nonuniform quantization* can provide fine quantization of the weak signals and coarse quantization of the strong signals. Thus in the case of nonuniform quantization, quantization noise can be made proportional to signal size. The effect is to improve the overall SNR by reducing the noise for the predominant weak signals, at the expense of an increase in noise for the rarely occurring strong signals. Figure 2.18 compares the quantization of a strong versus a weak signal for uniform and nonuniform quantization. The staircase-like waveforms represent the approximations to the analog waveforms (after quantization distortion has been introduced). The SNR improvement that nonuniform quantization provides for the weak signal should be apparent. Nonuniform quantization can be used to make the SNR a constant for all signals within the input range. For voice signals, the typical input signal

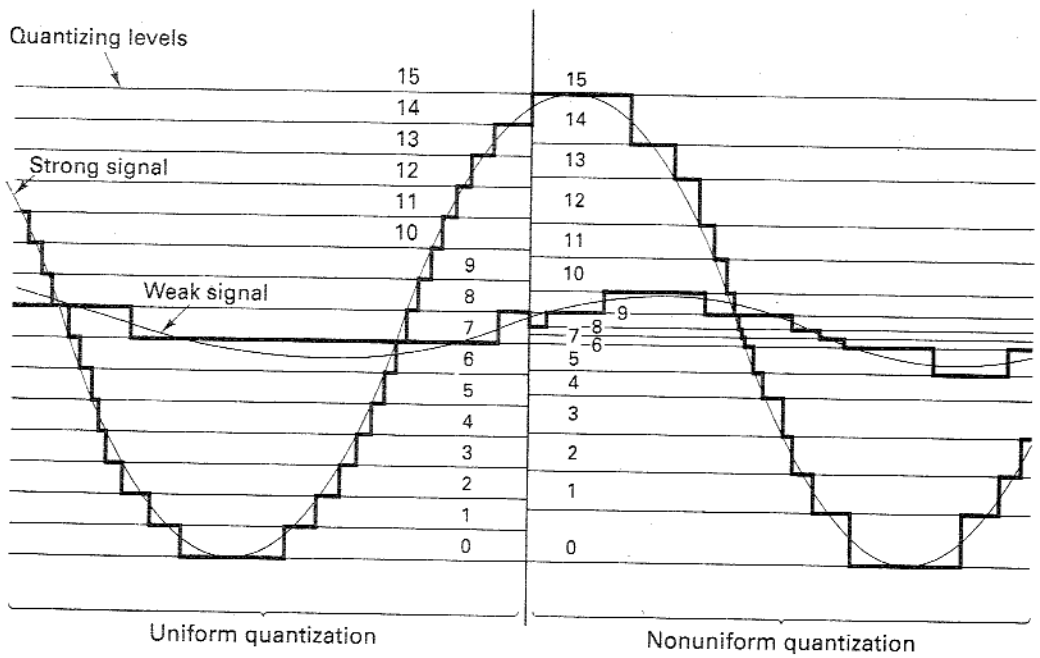

Figure 2.18 Uniform and nonuniform quantization of signals.

dynamic range is 40 decibels (dB), where a decibel is defined in terms of the ratio of power P_2 to power P_1:

$$\text{number of dB} = 10 \log_{10} \frac{P_2}{P_1} \tag{2.21}$$

With a uniform quantizer, weak signals would experience a 40-dB-poorer SNR than that of strong signals. The standard telephone technique of handling the large range of possible input signal levels is to use a *logarithmic-compressed* quantizer instead of a uniform one. With such a nonuniform compressor the output SNR is independent of the distribution of input signal levels.

2.7.2 Nonuniform Quantization

One way of achieving nonuniform quantization is to use a nonuniform quantizer characteristic, shown in Figure 2.19a. More often, nonuniform quantization is achieved by first distorting the original signal with a logarithmic compression characteristic, as shown in Figure 2.19b, and then using a uniform quantizer. For small magnitude signals the compression characteristic has a much steeper slope than for large magnitude signals. Thus, a given signal change at small magnitudes will carry the uniform quantizer through more steps than the same change at large magnitudes. The compression characteristic effectively changes the distribution of the

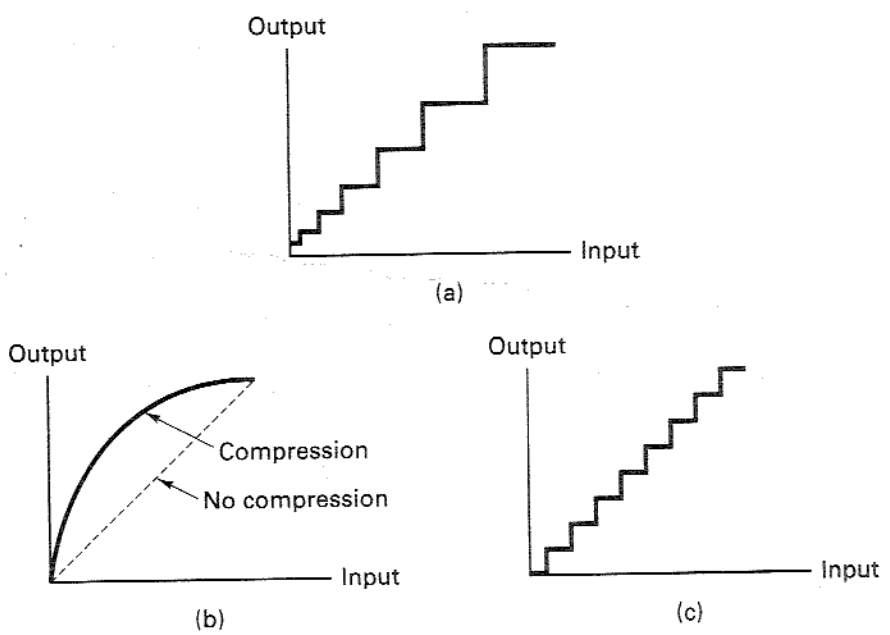

Figure 2.19 (a) Nonuniform quantizer characteristic. (b) Compression characteristic. (c) Uniform quantizer characteristic.

input signal magnitudes so that there is not a preponderance of *low* magnitude signals at the output of the compressor. After compression, the distorted signal is used as the input to a uniform (linear) quantizer characteristic, shown in Figure 2.19c. At the receiver, an inverse compression characteristic, called *expansion,* is applied so that the overall transmission is not distorted. The processing pair (compression and expansion) is usually referred to as *companding.*

2.7.3 Companding Characteristics

The early PCM systems implemented a smooth logarithmic compression function. Today, most PCM systems use a piecewise linear approximation to the logarithmic compression characteristic. In North America, a μ-law compression characteristic

$$y = y_{\max} \frac{\log_e[1 + \mu(|x| / x_{\max})]}{\log_e(1 + \mu)} \operatorname{sgn} x \tag{2.22}$$

is used, where

$$\operatorname{sgn} x = \begin{cases} +1 & \text{for } x \geq 0 \\ -1 & \text{for } x < 0 \end{cases}$$

and where μ is a positive constant, x and y represent input and output voltages, and $x_{\max}$ and $y_{\max}$ are the maximum positive excursions of the input and output voltages, respectively. The compression characteristic is shown in Figure 2.20a for several values of μ. In North America, the standard value for μ is 255. Notice that $\mu = 0$ corresponds to linear amplification (uniform quantization).

Another compression characteristic, used mainly in Europe, is the A-law characteristic, defined as

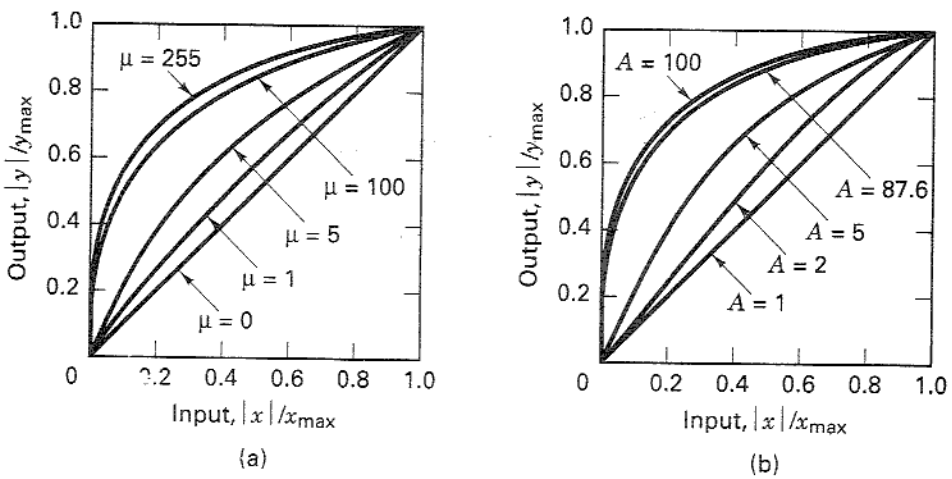

Figure 2.20 Compression characteristics. (a) μ-law characteristic. (b) A-law characteristic.

$$y = \begin{cases} y_{\max} \dfrac{A(|x|/x_{\max})}{1 + \log_e A} \operatorname{sgn} x & 0 < \dfrac{|x|}{x_{\max}} \le \dfrac{1}{A} \\ y_{\max} \dfrac{1 + \log_e[A(|x|/x_{\max})]}{1 + \log_e A} \operatorname{sgn} x & \dfrac{1}{A} < \dfrac{|x|}{x_{\max}} < 1 \end{cases} \tag{2.23}$$

where A is a positive constant and x and y are as defined in Equation (2.22). The A-law compression characteristic is shown in Figure 2.20b for several values of A. A standard value for A is 87.6. (The subjects of uniform and nonuniform quantization are treated further in Chapter 13, Section 13.2.)

2.8 BASEBAND TRANSMISSION

2.8.1 Waveform Representation of Binary Digits

In Section 2.6, it was shown how analog waveforms are transformed into binary digits via the use of PCM. There is nothing "physical" about the digits resulting from this process. Digits are just abstractions—a way to describe the message information. Thus, we need something physical that will represent or "carry" the digits.

We will represent the binary digits with electrical pulses in order to transmit them through a baseband channel. Such a representation is shown in Figure 2.21. Codeword time slots are shown in Figure 2.21a, where the codeword is a 4-bit representation of each quantized sample. In Figure 2.21b, each binary one is represented by a pulse and each binary zero is represented by the absence of a pulse. Thus a sequence of electrical pulses having the pattern shown in Figure 2.21b can be used to transmit the information in the PCM bit stream, and hence the information in the quantized samples of a message.

At the receiver, a determination must be made as to the presence or absence of a pulse in each bit time slot. It will be shown in Section 2.9 that the likelihood of correctly detecting the presence of a pulse is a function of the received pulse energy (or area under the pulse). Thus there is an advantage in making the pulse width T' in Figure 2.21b as wide as possible. If we increase the pulse width to the maximum possible (equal to the bit time T), we have the waveform shown in Figure 2.21c. Rather than describe this waveform as a sequence of present or absent pulses, we can describe it as a sequence of transitions between two levels. When the waveform occupies the upper voltage level it represents a binary one; when it occupies the lower voltage level it represents a binary zero.

2.8.2 PCM Waveform Types

When pulse modulation is applied to a *binary* symbol, the resulting binary waveform is called a pulse-code modulation (PCM) waveform. There are several types of PCM waveforms that are described below and illustrated in Figure 2.22; in telephony applications, these waveforms are often called *line codes*. When pulse modulation is applied to a *nonbinary* symbol, the resulting waveform is called an M-ary

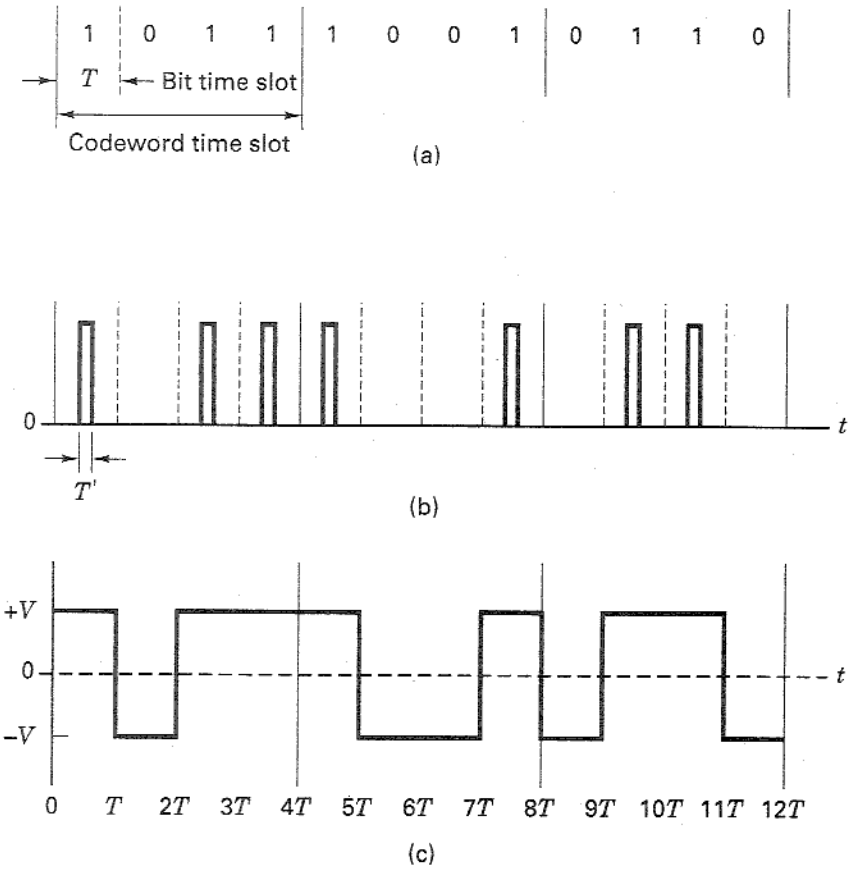

Figure 2.21 Example of waveform representation of binary digits. (a) PCM sequence. (b) Pulse representation of PCM. (c) Pulse waveform (transition between two levels).

pulse-modulation waveform, of which there are several types. They are described in Section 2.8.5, where one of them, called pulse-amplitude modulation (PAM), is emphasized. In Figure 2.1, the highlighted block, labeled *Baseband Signaling,* shows the basic classification of the PCM waveforms and the M-ary pulse waveforms. The PCM waveforms fall into the following four groups.

1. Nonreturn-to-zero (NRZ)
2. Return-to-zero (RZ)
3. Phase encoded
4. Multilevel binary

The NRZ group is probably the most commonly used PCM waveform. It can be partitioned into the following subgroups: NRZ-L (L for level), NRZ-M (M for mark), and NRZ-S (S for space). NRZ-L is used extensively in digital logic circuits.

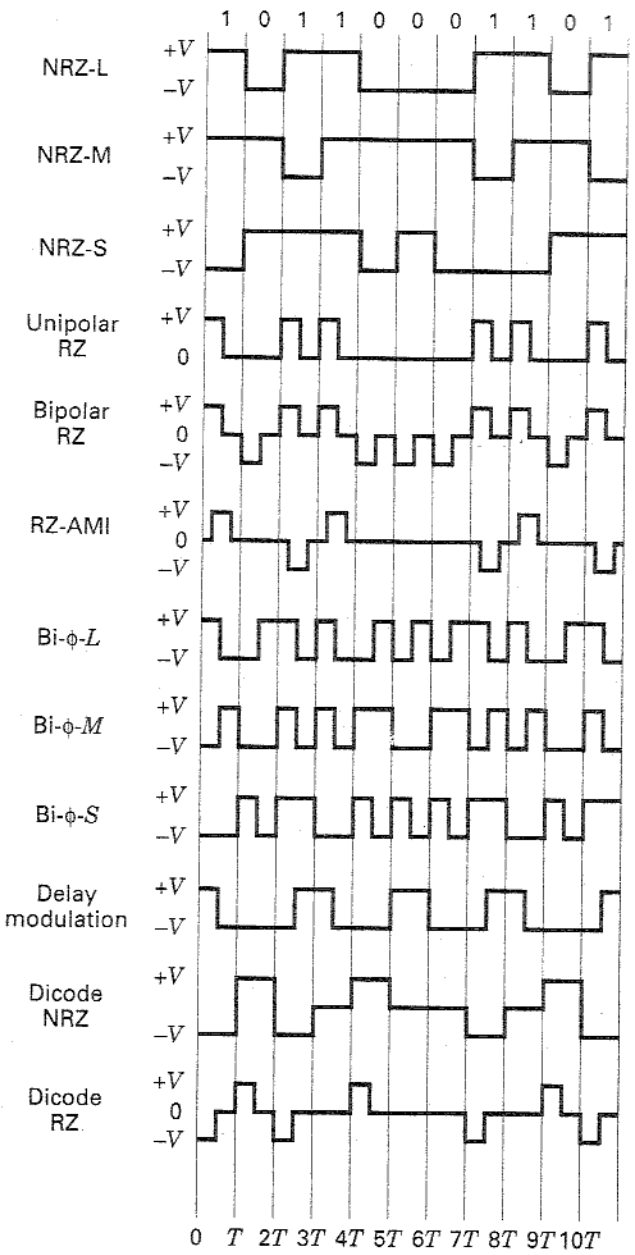

Figure 2.22 Various PCM waveforms.

A binary one is represented by one voltage level and a binary zero is represented by another voltage level. There is a change in level whenever the data change from a one to a zero or from a zero to a one. With NRZ-M, the one, or *mark*, is represented by a change in level, and the zero, or *space*, is represented by no change in level. This is often referred to as *differential encoding*. NRZ-M is used primarily in

magnetic tape recording. NRZ-S is the complement of NRZ-M: A one is represented by no change in level, and a zero is represented by a change in level.

The RZ waveforms consist of unipolar-RZ, bipolar-RZ, and RZ-AMI. These codes find application in baseband data transmission and in magnetic recording. With unipolar-RZ, a one is represented by a half-bit-wide pulse, and a zero is represented by the absence of a pulse. With bipolar-RZ, the ones and zeros are represented by opposite-level pulses that are one-half bit wide. There is a pulse present in each bit interval. RZ-AMI (AMI for "alternate mark inversion") is a signaling scheme used in telephone systems. The ones are represented by equal-amplitude alternating pulses. The zeros are represented by the absence of pulses.

The phase-encoded group consists of bi-ϕ-L (bi-phase-level), better known as *Manchester coding;* bi-ϕ-M (bi-phase-mark); bi-ϕ-S (bi-phase-space); and *delay modulation* (DM), or *Miller coding.* The phase-encoding schemes are used in magnetic recording systems and optical communications and in some satellite telemetry links. With bi-ϕ-L, a one is represented by a half-bit-wide pulse positioned during the first half of the bit interval; a zero is represented by a half-bit-wide pulse positioned during the second half of the bit interval. With bi-ϕ-M, a transition occurs at the beginning of every bit interval. A one is represented by a second transition one-half bit interval later; a zero is represented by no second transition. With bi-ϕ-S, a transition also occurs at the beginning of every bit interval. A one is represented by no second transition; a zero is represented by a second transition one-half bit interval later. With delay modulation [4], a one is represented by a transition at the midpoint of the bit interval. A zero is represented by no transition, unless it is followed by another zero. In this case, a transition is placed at the end of the bit interval of the first zero. Reference to the illustration in Figure 2.22 should help to make these descriptions clear.

Many binary waveforms use three levels, instead of two, to encode the binary data. Bipolar RZ and RZ-AMI belong to this group. The group also contains formats called *dicode* and *duobinary.* With dicode-NRZ, the one-to-zero or zero-to-one data transition changes the pulse polarity; without a data transition, the zero level is sent. With dicode-RZ, the one-to-zero or zero-to-one transition produces a half-duration polarity change; otherwise, a zero level is sent. The three-level duobinary signaling scheme is treated in Section 2.9.

One might ask why there are so many PCM waveforms. Are there really so many unique applications necessitating such a variety of waveforms to represent digits? The reason for the large selection relates to the differences in performance that characterize each waveform [5]. In choosing a PCM waveform for a particular application, some of the parameters worth examining are the following:

1. *Dc component.* Eliminating the dc energy from the signal's power spectrum enables the system to be ac coupled. Magnetic recording systems, or systems using transformer coupling, have little sensitivity to very low frequency signal components. Thus low-frequency information could be lost.
2. *Self-Clocking.* Symbol or bit synchronization is required for any digital communication system. Some PCM coding schemes have inherent synchronizing

or clocking features that aid in the recovery of the clock signal. For example, the Manchester code has a transition in the middle of every bit interval whether a one or a zero is being sent. This guaranteed transition provides a clocking signal.

3. *Error detection.* Some schemes, such as duobinary, provide the means of detecting data errors without introducing additional error-detection bits into the data sequence.
4. *Bandwidth compression.* Some schemes, such as multilevel codes, increase the efficiency of bandwidth utilization by allowing a reduction in required bandwidth for a given data rate; thus there is more information transmitted per unit bandwidth.
5. *Differential encoding.* This technique is useful because it allows the polarity of differentially encoded waveforms to be inverted without affecting the data detection. In communication systems where waveforms sometimes experience inversion, this is a great advantage. (Differential encoding is treated in greater detail in Chapter 4, Section 4.5.2.)
6. *Noise immunity.* The various PCM waveform types can be further characterized by probability of bit error versus signal-to-noise ratio. Some of the schemes are more immune than others to noise. For example, the NRZ waveforms have better error performance than does the unipolar RZ waveform.

2.8.3 Spectral Attributes of PCM Waveforms

The most common criteria used for comparing PCM waveforms and for selecting one waveform type from the many available are spectral characteristics, bit synchronization capabilities, error-detecting capabilities, interference and noise immunity, and cost and complexity of implementation. Figure 2.23 shows the spectral characteristics of some of the most popular PCM waveforms. The figure plots power spectral density in watts/hertz versus normalized bandwidth, WT, where W is bandwidth, and T is the duration of the pulse. WT is often referred to as the *time-bandwidth product,* of the signal. Since the pulse or symbol rate R_s is the reciprocal of T, normalized bandwidth can also be expressed as W/R_s. From this latter expression, we see that the units of normalized bandwidth are hertz/(pulse/s) or hertz/(symbol/s). This is a relative measure of bandwidth; it is valuable because it describes how efficiently the transmission bandwidth is being utilized for each waveform of interest. Any waveform type that requires less than 1.0 Hz for sending 1 symbol/s is relatively bandwidth efficient. Examples would be delay modulation and duobinary (see Section 2.9). By comparison, any waveform type that requires more than 1.0 Hz for sending 1 symbol/s is relatively bandwidth inefficient. An example of this would be bi-phase (Manchester) signaling. From Figure 2.23, we can also see the spectral concentration of signaling energy for each waveform type. For example, NRZ and duobinary schemes have large spectral components at dc and low frequency, while bi-phase has no energy at dc.

An important parameter for measuring *bandwidth efficiency* is R/W having units of bits/s/hz. This measure involves data rate rather than symbol rate. For a

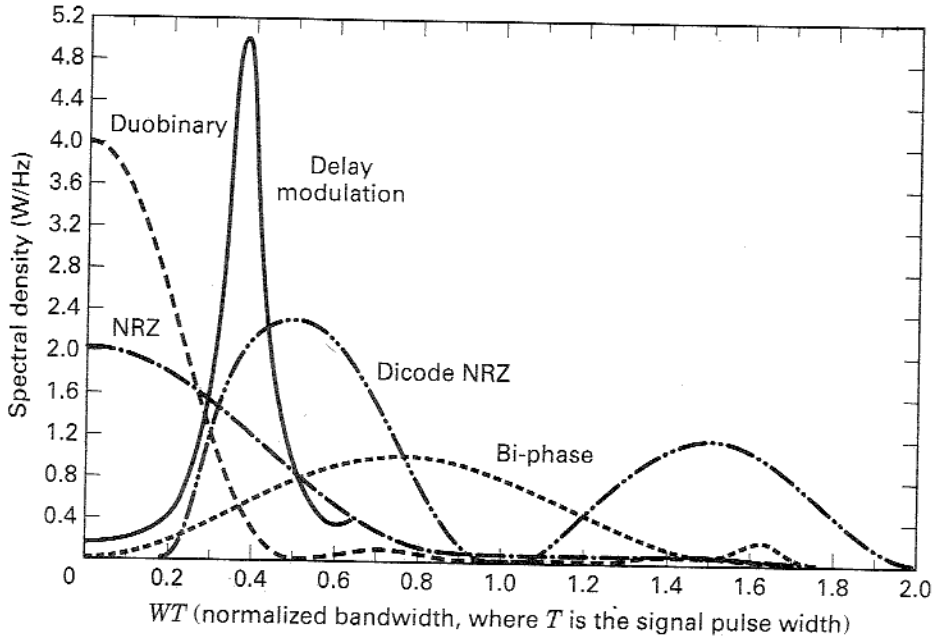

Figure 2.23 Spectral densities of various PCM waveforms.

given signaling scheme, R/W describes how much data throughput can be transmitted for each Hertz of available bandwidth. (Bandwidth efficiency is treated in greater detail in Chapter 9.)

2.8.4 Bits per PCM Word and Bits per Symbol

Throughout Chapters 1 and 2, the idea of binary partitioning ($M = 2^k$) is used to relate the grouping of bits to form symbols for the purpose of signal processing and transmission. We now examine an analogous application where the $M = 2^k$ concept is also applicable. Consider the process of formatting analog information into a bit steam via sampling, quantization, and coding. Each analog sample is transformed into a PCM word made up of groups of bits. The PCM word size can be described by the number of quantization levels allowed for each sample; this is identical to the number of values that the PCM word can assume. Or, the quantization can be described by the number of bits required to identify that set of levels. The relationship between the number of levels per sample and the number of bits needed to represent those levels is the same as the $M = 2^k$ relationship between the size of a set of message symbols and the number of bits needed to represent the symbol. To distinguish between the two applications, the notation is changed for the PCM case. Instead of $M = 2^k$, we use $L = 2^\ell$, where L is the number of quantization levels in the PCM word, and ℓ is the number of bits needed to represent those levels.

2.8.4.1 PCM Word Size

How many bits shall we assign to each analog sample? For digital telephone channels, each speech sample is PCM encoded using 8 bits, yielding 2^8 or 256 levels per sample. The choice of the number of levels, or bits per sample, depends on how much quantization distortion we are willing to tolerate with the PCM format. It is useful to develop a general relationship between the required number of bits per analog sample (the PCM word size), and the allowable quantization distortion. Let the magnitude of the quantization distortion error, $|e|$, be specified as a fraction p of the peak-to-peak analog voltage V_{pp} as follows:

$$|e| \leq pV_{pp} \tag{2.24}$$

Since the quantization error can be no larger than $q/2$, where q is the quantile interval, we can write

$$|e|_{max} = \frac{q}{2} = \frac{V_{pp}}{2(L-1)} \approx \frac{V_{pp}}{2L} \tag{2.25}$$

where L is the number of quantization levels. For most applications the number of levels is large enough so that $L-1$ can be replaced by L, as was done above. Then, from Equations (2.24) and (2.25), we can write

$$\frac{V_{pp}}{2L} \leq pV_{pp} \tag{2.26}$$

$$2^{\ell} = L \geq \frac{1}{2p} \quad \text{levels} \tag{2.27}$$

and

$$\ell \geq \log_2 \frac{1}{2p} \quad \text{bits} \tag{2.28}$$

It is important that we do not confuse the idea of bits per PCM word, denoted by ℓ in Equation (2.28), with the M-level transmission concept of k data bits per symbol. (Example 2.3, presented shortly, should clarify the distinction.)

2.8.5 *M*-ary Pulse-Modulation Waveforms

There are three basic ways to modulate information on to a sequence of pulses: we can vary the pulse's amplitude, position, or duration, which leads to the names *pulse-amplitude modulation* (PAM), *pulse-position modulation* (PPM), and *pulse-duration modulation* (PDM), respectively. PDM is sometimes called pulse-width modulation (PWM). When information samples without any quantization are modulated on to pulses, the resulting pulse modulation can be called *analog pulse modulation.* When the information samples are first quantized, yielding symbols from an M-ary alphabet set, and then modulated on to pulses, the resulting pulse modulation is digital and we refer to it as *M-ary pulse modulation.* In the case of M-ary PAM, one of M allowable amplitude levels are assigned to each of the M possible symbol values. Earlier we described PCM waveforms as binary waveforms having

two amplitude values (e.g., NRZ, RZ). Note that such PCM waveforms requiring only two levels represent the special case ($M = 2$) of the general M-ary PAM that requires M levels. In this book, the PCM waveforms are grouped separately (see Figure 2.1 and Section 2.8.2) and are emphasized because they are the most popular of the pulse-modulation schemes.

In the case of M-ary PPM waveforms, modulation is effected by delaying (or advancing) a pulse occurrence, by an amount that corresponds to the value of the information symbols. In the case of M-ary PDM waveforms, modulation is effected by varying the pulse width by an amount that corresponds to the value of the symbols. For both PPM and PDM, the pulse amplitude is held constant. Baseband modulation with pulses have analogous counterparts in the area of bandpass modulation. PAM is similar to amplitude modulation, while PPM and PDM are similar to phase and frequency modulation respectively. In this section, we only address M-ary PAM waveforms as they compare to PCM waveforms.

The transmission bandwidth required for binary digital waveforms such as PCM may be very large. What might we do to reduce the required bandwidth? One possibility is to use *multilevel signaling*. Consider a bit stream with data rate, R bits per second. Instead of transmitting a pulse waveform for each bit, we might first partition the data into k-bit groups, and then use ($M = 2^k$)-level pulses for transmission. With such multilevel signaling or M-ary PAM, each pulse waveform can now represent a k-bit symbol in a symbol stream moving at the rate of R/k symbols per second (a factor k slower than the bit stream). Thus for a given data rate, multilevel signaling, where $M > 2$, can be used to reduce the number of symbols transmitted per second; or, in other words, M-ary PAM as opposed to binary PCM can be used to reduce the transmission bandwidth requirements of the channel. Is there a price to be paid for such bandwidth reduction? Of course, and that is discussed below.

Consider the task that the pulse receiver must perform: It must distinguish between the possible levels of each pulse. Can the receiver distinguish among the eight possible levels of each octal pulse in Figure 2.24a as easily as it can distinguish between the two possible levels of each binary pulse in Figure 2.24b? The transmission of an 8-level (compared with a 2-level) pulse requires a greater amount of energy for equivalent detection performance. (It is the amount of received E_b/N_0 that determines how reliably a signal will be detected). For equal average power in the binary and the octal pulses, it is easier to detect the binary pulses because the detector has more signal energy per level for making a binary decision that an 8-level decision. What price does a system designer pay if he or she chooses the transmission waveform to be the easier-to-detect binary PCM rather than the 8-level PAM? The engineer pays the price of needing three times as much transmission bandwidth for a given data rate, compared with the octal pulses, since each octal pulse must be replaced with three binary pulses (each one-third as wide as the octal pulses). One might ask, Why not use binary pulses with the same pulse duration as the original octal pulses and suffer the information delay? For some cases, this might be appropriate, but for real-time communication systems, such an increase in delay cannot be tolerated—the 6 o'clock news *must* be received at 6 o-clock. (In Chapter 9, we examine in detail the trade-off between signal power and transmission bandwidth.)

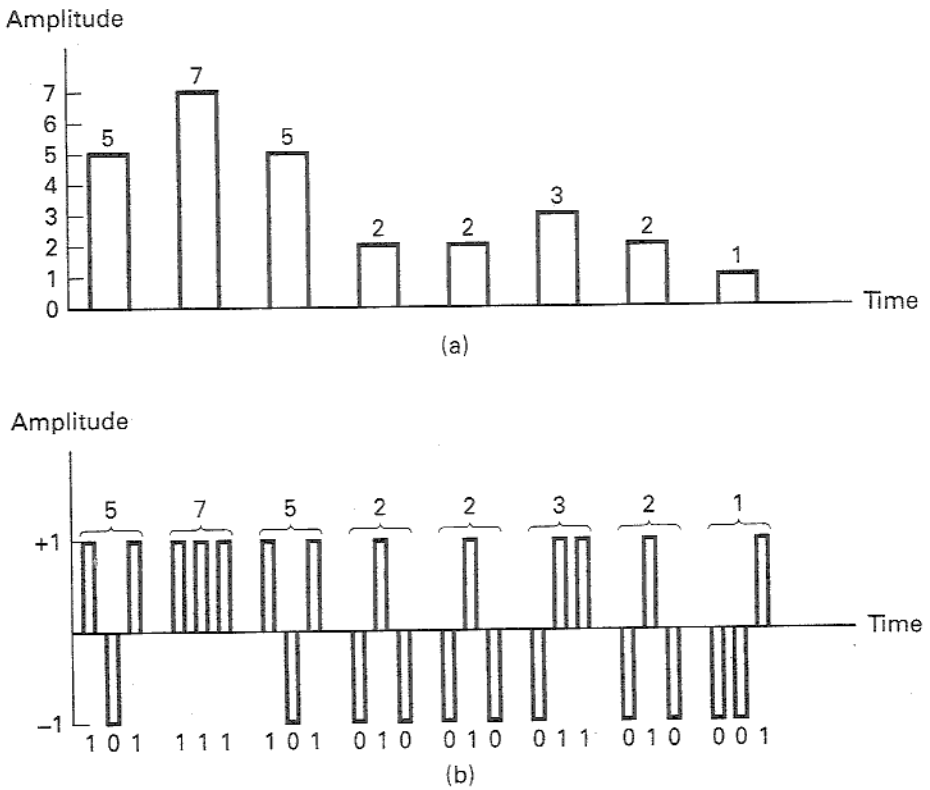

Figure 2.24 Pulse code modulation signaling. (a) Eight-level signaling. (b) Two-level signaling.

Example 2.3 Quantization Levels and Multilevel Signaling

The information in an analog waveform, with maximum frequency $f_m = 3$ kHz, is to be transmitted over an M-ary PAM system, where the number of pulse levels is $M = 16$. The quantization distortion is specified not to exceed ±1% of the peak-to-peak analog signal.

(a) What is the minimum number of bits/sample, or bits/PCM word that should be used in digitizing the analog waveform?
(b) What is the minimum required sampling rate, and what is the resulting bit transmission rate?
(c) What is the PAM pulse or symbol transmission rate?
(d) If the transmission bandwidth (including filtering) equals 12 kHz, determine the bandwidth efficiency for this system.

In this example we are concerned with two types of *levels:* the number of quantization levels for fulfilling the distortion requirement and the 16 levels of the multilevel PAM pulses.

Solution

(a) Using Equation (2.28), we calculate

$$\ell \geq \log_2 \frac{1}{0.02} = \log_2 50 \approx 5.6.$$

Therefore, use $\ell = 6$ bits/sample to meet the distortion requirement.

(b) Using the Nyquist sampling criterion, the minimum sampling rate $f_s = 2f_m = 6000$ samples/second. From part (a), each sample will give rise to a PCM word composed of 6 bits. Therefore the bit transmission rare $R = \ell f_s = 36{,}000$ bits/sec.

(c) Since multilevel pulses are to be used with $M = 2^k = 16$ levels, then $k = \log_2 16 = 4$ bits/symbol. Therefore, the bit stream will be partitioned into groups of 4 bits to form the new 16-level PAM digits, and the resulting symbol transmission rate R_s is $R/k = 36{,}000/4 = 9000$ symbols/s.

(d) Bandwidth efficiency is described by data throughput per hertz, R/W. Since $R = 36{,}000$ bits/s, and $W = 12$ kHz, then $R/W = 3$ bits/s/hz.

2.9 CORRELATIVE CODING

In 1963, Adam Lender [6, 7] showed that it is possible to transmit $2W$ symbols/s with zero ISI, using the theoretical minimum bandwidth of W hertz, without infinitely sharp filters. Lender used a technique called *duobinary signaling*, also referred to as *correlative coding* and *partial response signaling*. The basic idea behind the duobinary technique is to introduce some controlled amount of ISI into the data stream rather than trying to eliminate it completely. By introducing correlated interference between the pulses, and by changing the detection procedure, Lender, in effect, "canceled out" the interference at the detector and thereby achieved the ideal symbol-rate packing of 2 symbols/s/Hz, an amount that had been considered unrealizable.

2.9.1 Duobinary Signaling

To understand how duobinary signaling introduces controlled ISI, let us look at a model of the process. We can think of the duobinary coding operation as if it were implemented as shown in Figure 2.25. Assume that a sequence of binary symbols $\{x_k\}$ is to be transmitted at the rate of R symbols/s over a system having an ideal rectangular spectrum of bandwidth $W = R/2 = 1/2T$ hertz. You might ask: How is this rectangular spectrum, in Figure 2.25, different from the unrealizable Nyquist characteristic? It has the same ideal characteristic; but we are not trying to implement the ideal rectangular filter. It is only the part of our equivalent model that is used for developing a filter that is easier to approximate. Before being shaped by the ideal filter, the pulses pass through a simple digital filter, as shown in the figure. The digital filter incorporates a one-digit delay; to each incoming pulse, the filter adds the value of the previous pulse. In other words, for every pulse into the digital filter, we get the summation of two pulses out. Each pulse of the sequence $\{y_k\}$ out of the digital filter can be expressed as

$$y_k = x_k + x_{k-1} \tag{2.29}$$

a spectral shift in the signal. As a result of this equivalence theorem, all linear signal-processing simulations can take place at baseband (which is preferred for simplicity) with the same results as at bandpass. This means that the performance of most digital communication systems will often be described and analyzed as if the transmission channel is a baseband channel.

3.1 SIGNALS AND NOISE

3.1.1 Error-Performance Degradation in Communication Systems

The task of the detector is to retrieve the bit stream from the received waveform, as error free as possible, notwithstanding the impairments to which the signal may have been subjected. There are two primary causes for error-performance degradation. The first is the effect of filtering at the transmitter, channel, and receiver, discussed in Section 3.3, below. As described there, a nonideal system transfer function causes symbol "smearing" or *intersymbol interference* (ISI).

Another cause for error-performance degradation is electrical noise and interference produced by a variety of sources, such as galaxy and atmospheric noise, switching transients, intermodulation noise, as well as interfering signals from other sources. (These are discussed in Chapter 5.) With proper precautions, much of the noise and interference entering a receiver can be reduced in intensity or even eliminated. However, there is one noise source that cannot be eliminated, and that is the noise caused by the thermal motion of electrons in any conducting media. This motion produces *thermal noise* in amplifiers and circuits, and corrupts the signal in an additive fashion. The statistics of thermal noise have been developed using quantum mechanics, and are well known [1].

The primary statistical characteristic of thermal noise is that the noise amplitudes are distributed according to a normal or Gaussian distribution, discussed in Section 1.5.5, and shown in Figure 1.7. In this figure, it can be seen that the most probable noise amplitudes are those with small positive or negative values. In theory, the noise can be infinitely large, but very large noise amplitudes are rare. The primary spectral characteristic of thermal noise in communication systems, is that its two-sided power spectral density $G_n(f) = N_0/2$ is flat for all frequencies of interest. In other words, the thermal noise, on the average, has just as much power per hertz in low-frequency fluctuations as in high-frequency fluctuations—up to a frequency of about 10^{12} hertz. When the noise power is characterized by such a constant-power spectral density, we refer to it as *white noise.* Since thermal noise is present in all communication systems and is the predominant noise source for many systems, the thermal noise characteristics (additive, white, and Gaussian, giving rise to the name AWGN) are most often used to model the noise in the detection process and in the design of receivers. Whenever a channel is designated as an AWGN channel (with no other impairments specified), we are in effect being told that its impairments are limited to the degradation caused by this unavoidable thermal noise.

3.1.2 Demodulation and Detection

During a given signaling interval T, a binary baseband system will transmit one of two waveforms, denoted $g_1(t)$ and $g_2(t)$. Similarly, a binary bandpass system will transmit one of two waveforms, denoted $s_1(t)$ and $s_2(t)$. Since the general treatment of demodulation and detection are essentially the same for baseband and bandpass systems, we use $s_i(t)$ here as a generic designation for a transmitted waveform, whether the system is baseband or bandpass. This allows much of the baseband demodulation/detection treatment in this chapter to be consistent with similar bandpass descriptions in Chapter 4. Then, for any binary channel, the transmitted signal over a symbol interval $(0, T)$ is represented by

$$s_i(t) = \begin{cases} s_1(t) & 0 \le t \le T \quad \text{for a binary 1} \\ s_2(t) & 0 \le t \le T \quad \text{for a binary 0} \end{cases}$$

The received signal $r(t)$ degraded by noise $n(t)$ and possibly degraded by the impulse response of the channel $h_c(t)$ was described in Equation (1.1) and is rewritten as

$$r(t) = s_i(t) * h_c(t) + n(t) \qquad i = 1, \ldots, M \tag{3.1}$$

where $n(t)$ is here assumed to be a zero mean AWGN process, and $*$ represents a convolution operation. For binary transmission over an ideal distortionless channel where convolution with $h_c(t)$ produces no degradation (since for the ideal case $h_c(t)$ is an impulse function), the representation of $r(t)$ can be simplified to

$$r(t) = s_i(t) + n(t) \qquad i = 1, 2, \qquad 0 \le t \le T \tag{3.2}$$

Figure 3.1 shows the typical demodulation and detection functions of a digital receiver. Some authors use the terms "demodulation" and "detection" interchangeably. This book makes a distinction between the two. We define *demodulation* as recovery of a waveform (to an undistorted baseband pulse), and we designate *detection* to mean the decision-making process of selecting the digital meaning of that waveform. If error-correction coding *not* present, the detector output consists of estimates of message symbols (or bits), $\hat{m}_i$ (also called *hard decisions*). If error-correction coding is used, the detector output consists of estimates of channel symbols (or coded bits) $\hat{u}_i$, which can take the form of *hard* or *soft decisions* (see Section 7.3.2). For brevity, the term "detection" is occasionally used loosely to encompass all the receiver signal-processing steps through the decision making step. The *frequency down-conversion* block, shown in the demodulator portion of Figure 3.1, performs frequency translation for bandpass signals operating at some radio frequency (RF). This function may be configured in a variety of ways. It may take place within the front end of the receiver, within the demodulator, shared between the two locations, or not at all.

Within the *demodulate* and *sample* block of Figure 3.1 is the *receiving filter* (essentially the demodulator), which performs waveform recovery in preparation for the next important step—detection. The filtering at the transmitter and the channel typically cause the received pulse sequence to suffer from ISI, and thus it is

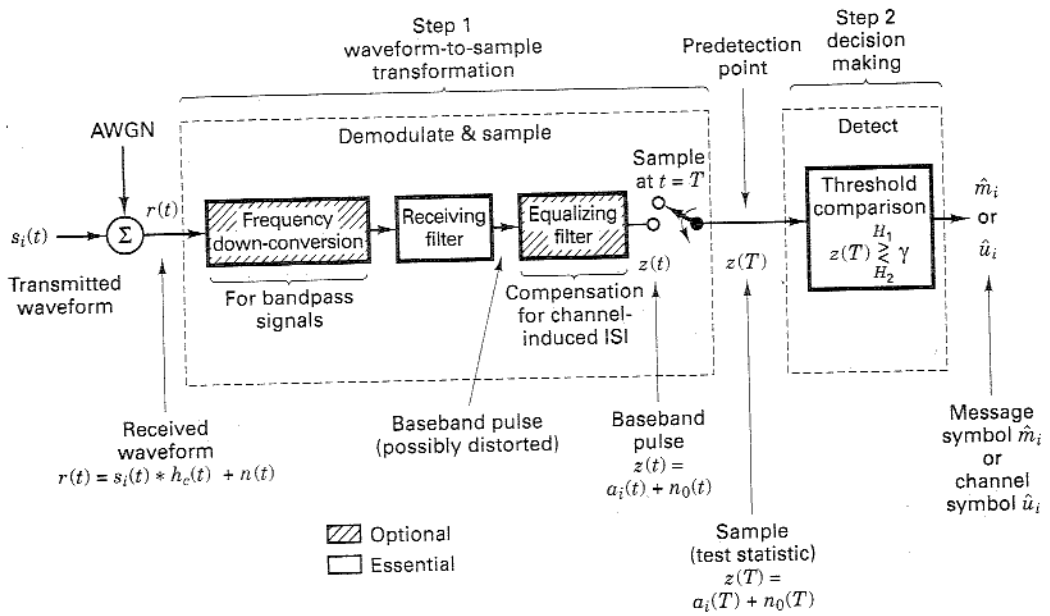

Figure 3.1 Two basic steps in the demodulation/detection of digital signals.

not quite ready for sampling and detection. The goal of the receiving filter is to recover a baseband pulse with the best possible signal-to-noise ratio (SNR), free of any ISI. The optimum receiving filter for accomplishing this is called a *matched filter* or *correlator,* described in Sections 3.2.2 and 3.2.3. An optional *equalizing filter* follows the receiving filter; it is only needed for those systems where channel-induced ISI can distort the signals. The receiving filter and equalizing filter are shown as two separate blocks in order to emphasize their separate functions. In most cases, however, when an equalizer is used, a single filter would be designed to incorporate both functions and thereby compensate for the distortion caused by both the transmitter and the channel. Such a composite filter is sometimes referred to simply as the *equalizing filter* or the *receiving and equalizing filter.*

Figure 3.1 highlights two steps in the demodulation/detection process. Step 1, the waveform-to-sample transformation, is made up of the demodulator followed by a sampler. At the end of each symbol duration T, the output of the sampler, the *predetection point,* yields a sample $z(T)$, sometimes called the test statistic. $z(T)$ has a voltage value directly proportional to the energy of the received symbol and inversely proportional to the noise. In step 2, a decision (detection) is made regarding the digital meaning of that sample. We assume that the input noise is a Gaussian random process and that the receiving filter in the demodulator is linear. A linear operation performed on a Gaussian random process will produce a second Gaussian random process [2]. Thus, the filter output noise is Gaussian. The output of step 1 yields the test statistic

$$z(T) = a_i(T) + n_0(T) \qquad i = 1, 2 \tag{3.3}$$

where $a_i(T)$ is the desired signal component, and $n_0(T)$ is the noise component. To simplify the notation, we sometimes express Equation (3.3) in the form of $z = a_i + n_0$. The noise component n_0 is a zero mean Gaussian random variable, and thus $z(T)$ is a Gaussian random variable with a mean of either a_1 or a_2 depending on whether a binary one or binary zero was sent. As described in Section 1.5.5, the probability density function (pdf) of the Gaussian random noise n_0 can be expressed as

$$p(n_0) = \frac{1}{\sigma_0 \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{n_0}{\sigma_0}\right)^2\right] \tag{3.4}$$

where σ_0^2 is the noise variance. Thus it follows from Equations (3.3) and (3.4) that the conditional pdfs $p(z|s_1)$ and $p(z|s_2)$ can be expressed as

$$p(z|s_1) = \frac{1}{\sigma_0 \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{z - a_1}{\sigma_0}\right)^2\right] \tag{3.5}$$

and

$$p(z|s_2) = \frac{1}{\sigma_0 \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{z - a_2}{\sigma_0}\right)^2\right] \tag{3.6}$$

These conditional pdfs are illustrated in Figure 3.2. The rightmost conditional pdf, $p(z|s_1)$, called the *likelihood* of s_1, illustrates the probability density function of the random variable $z(T)$, given that symbol s_1 was transmitted. Similarly, the leftmost conditional pdf, $p(z|s_2)$, called the *likelihood* of s_2, illustrates the pdf of $z(T)$, given that symbol s_2 was transmitted. The abscissa, $z(T)$, represents the full range of possible sample output values from step 1 of Figure 3.1.

After a received waveform has been transformed to a sample, the actual shape of the waveform is no longer important; all waveform types that are transformed to the same value of $z(T)$ are identical for detection purposes. Later it is shown that an optimum receiving filter (matched filter) in step 1 of Figure 3.1 maps all signals of equal energy into the same point $z(T)$. Therefore, the received *signal energy* (not its shape) is the important parameter in the detection process. This is why the detection analysis for baseband signals is the same as that for bandpass sig-

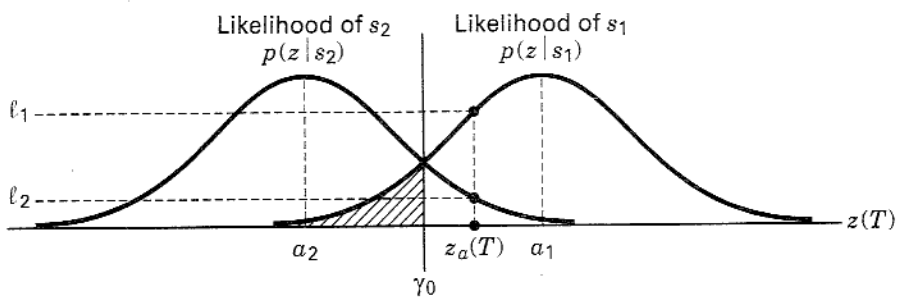

Figure 3.2 Conditional probability density functions: $p(z|s_1)$ and $p(z|s_2)$.

nals. Since $z(T)$ is a voltage signal that is proportional to the energy of the received symbol, the larger the magnitude of $z(T)$, the more error free will be the decision-making process. In step 2, detection is performed by choosing the hypothesis that results from the threshold measurement

$$z(T) \underset{H_2}{\overset{H_1}{\gtrless}} \gamma \tag{3.7}$$

where H_1 and H_2 are the two possible (binary) hypotheses. The inequality relationship indicates that hypothesis H_1 is chosen if $z(T) > \gamma$, and hypothesis H_2 is chosen if $z(T) < \gamma$. If $z(T) = \gamma$, the decision can be an arbitrary one. Choosing H_1 is equivalent to deciding that signal $s_1(t)$ was sent and hence a binary 1 is detected. Similarly, choosing H_2 is equivalent to deciding that signal $s_2(t)$ was sent, and hence a binary 0 is detected.

3.1.3 A Vector View of Signals and Noise

We now present a geometric or vector view of signal waveforms that are useful for either baseband or bandpass signals. We define an N-dimensional *orthogonal space* as a space characterized by a set of N linearly independent functions $\{\phi_j(t)\}$, called *basis functions.* Any arbitrary function in the space can be generated by a linear combination of these basis functions. The basis functions must satisfy the conditions

$$\int_0^T \psi_j(t)\psi_k(t)\,dt = K_j\delta_{jk} \qquad 0 \le t \le T \qquad j,k = 1,\ldots,N \tag{3.8a}$$

where the operator

$$\delta_{jk} = \begin{cases} 1 & \text{for } j = k \\ 0 & \text{otherwise} \end{cases} \tag{3.8b}$$

is called the *Kronecker delta function* and is defined by Equation (3.8b). When the K_j constants are nonzero, the signal space is called *orthogonal.* When the basis functions are normalized so that each $K_j = 1$, the space is called an *orthonormal* space. The principal requirement for orthogonality can be stated as follows. Each $\psi_j(t)$ function of the set of basis functions must be independent of the other members of the set. Each $\psi_j(t)$ must not interfere with any other members of the set in the detection process. From a geometric point of view, each $\psi_j(t)$ is mutually perpendicular to each of the other $\psi_k(t)$ for $j \neq k$. An example of such a space with $N = 3$ is shown in Figure 3.3, where the mutually perpendicular axes are designated $\psi_1(t)$, $\psi_2(t)$, and $\psi_3(t)$. If $\psi_j(t)$ corresponds to a real-valued voltage or current waveform component, associated with a 1-Ω resistive load, then using Equations (1.5) and (3.8), the normalized energy in joules dissipated in the load in T seconds, due to ψ_j, is

The fact that a digital signal is best characterized by its received energy doesn't yet get to the crux of why E_b/N_0 is a natural metric for digital systems, so let us continue. The digital waveform is a vehicle that represents a digital message. The message may contain one bit (binary), two bits (4-ary), . . . , 10 bits (1024-ary). In analog systems, there is nothing akin to such a discretized message structure. An analog information source is an infinitely quantized continuous wave. For digital systems, a figure of merit should allow us to compare one system with another at the bit level. Therefore, a description of the digital waveform in terms of *S/N* is virtually useless, since the waveform may have a one-bit meaning, a two-bit meaning, or a 10-bit meaning. For example, suppose we are told that for a given error probability, the required *S/N* for a digital binary waveform is 20 units. Think of the waveform as being interchangeable with its meaning. Since the binary waveform has a one-bit meaning, then the *S/N* requirement per bit is equal to the same 20 units. However, suppose that the waveform is 1024-ary, with the same 20 units of required *S/N*. Now, since the waveform has a 10-bit meaning, the *S/N* requirement per bit is only 2 units. Why should we have to go through such computational manipulations to find a metric that represents a figure of merit? Why not immediately describe the metric in terms of what we need—an energy-related parameter at the bit level, E_b/N_0? Just as *S/N* is a dimensionless ratio, so too is E_b/N_0. To verify this, consider the following units of measure:

$$\frac{E_b}{N_0} = \frac{\text{Joule}}{\text{Watt per Hz}} = \frac{\text{Watt-s}}{\text{Watt-s}}$$

3.2 DETECTION OF BINARY SIGNALS IN GAUSSIAN NOISE

3.2.1 Maximum Likelihood Receiver Structure

The decision-making criterion shown in step 2 of Figure 3.1 was described by Equation (3.7) as

$$z(T) \underset{H_2}{\overset{H_1}{\gtrless}} \gamma$$

A popular criterion for choosing the threshold level γ for the binary decision in Equation (3.7) is based on minimizing the probability of error. The computation for this *minimum error* value of $\gamma = \gamma_0$ starts with forming an inequality expression between the ratio of conditional probability density functions and the signal a priori probabilities. Since the conditional density function $p(z|s_i)$ is also called the *likelihood* of s_i, the formulation

$$\frac{p(z|s_1)}{p(z|s_2)} \underset{H_2}{\overset{H_1}{\gtrless}} \frac{P(s_2)}{P(s_1)} \tag{3.31}$$

is called the *likelihood ratio test.* (See Appendix B.) In this inequality, $P(s_1)$ and $P(s_2)$ are the a priori probabilities that $s_1(t)$ and $s_2(t)$, respectively, are transmitted, and H_1 and H_2 are the two possible hypotheses. The rule for minimizing the error probability states that we should choose hypothesis H_1 if the ratio of likelihoods is greater than the ratio of a priori probabilities, as shown in Equation (3.31).

It is shown in Section B.3.1, that if $P(s_1) = P(s_2)$, and if the likelihoods, $p(z|s_i)$ $(i = 1, 2)$, are symmetrical, the substitution of Equations (3.5) and (3.6) into (3.31) yields

$$z(T) \underset{H_2}{\overset{H_1}{\gtrless}} \frac{a_1 + a_2}{2} = \gamma_0 \tag{3.32}$$

where a_1 is the signal component of $z(T)$ when $s_1(t)$ is transmitted, and a_2 is the signal component of $z(T)$ when $s_2(t)$ is transmitted. The threshold level γ_0, represented by $(a_1 + a_2)/2$, is the *optimum threshold* for minimizing the probability of making an incorrect decision for this important special case. This strategy is known as the *minimum error criterion.*

For equally likely signals, the optimum threshold γ_0 passes through the intersection of the likelihood functions, as shown in Figure 3.2. Thus by following Equation (3.32), the decision stage effectively selects the hypothesis that corresponds to the signal with the *maximum likelihood.* For example, given an arbitrary detector output value $z_a(T)$, for which there is a nonzero likelihood that $z_a(T)$ belongs to either signal class $s_1(t)$ or $s_2(t)$, one can think of the likelihood test as a comparison of the likelihood values $p(z_a|s_1)$ and $p(z_a|s_2)$. The signal corresponding to the maximum pdf is chosen as the most likely to have been transmitted. In other words, the detector chooses $s_1(t)$ if

$$p(z_a|s_1) > p(z_a|s_2) \tag{3.33}$$

Otherwise, the detector chooses $s_2(t)$. A detector that minimizes the error probability (for the case where the signal classes are equally likely) is also known as a *maximum likelihood detector.*

Figure 3.2 illustrates that Equation (3.33) is just a "common sense" way to make a decision when there exists statistical knowledge of the classes. Given the detector output value $z_a(T)$, we see in Figure 3.2 that $z_a(T)$ intersects the likelihood of $s_1(t)$ at a value ℓ_1, and it intersects the likelihood of $s_2(t)$ at a value ℓ_2. What is the most reasonable decision for the detector to make? For this example, choosing class $s_1(t)$, which has the greater likelihood, is the most sensible choice. If this was an M-ary instead of a binary example, there would be a total of M likelihood functions representing the M signal classes to which a received signal might belong. The maximum likelihood decision would then be to choose the class that had the greatest likelihood of all M likelihoods. (Refer to Appendix B for a review of decision theory fundamentals.)

3.2.1.1 Error Probability

For the binary decision-making depicted in Figure 3.2, there are two ways errors can occur. An error e will occur when $s_1(t)$ is sent, and channel noise results in the receiver output signal $z(t)$ being less than γ_0. The probability of such an occurrence is

$$P(e\,|s_1) = P(H_2|s_1) = \int_{-\infty}^{\gamma_0} p(z\,|s_1)\,dz \tag{3.34}$$

This is illustrated by the shaded area to the left of γ_0 in Figure 3.2. Similarly, an error occurs when $s_2(t)$ is sent, and the channel noise results in $z(T)$ being greater than γ_0. The probability of this occurrence is

$$P(e\,|s_2) = P(H_1|s_2) = \int_{\gamma_0}^{\infty} p(z\,|s_2)\,dz \tag{3.35}$$

The probability of an error is the sum of the probabilities of all the ways that an error can occur. For the binary case, we can express the probability of bit error as

$$P_B = \sum_{i=1}^{2} P(e, s_i) = \sum_{i=1}^{2} P(e\,|s_i)\,P(s_i) \tag{3.36}$$

Combining Equations (3.34) to (3.36), we can write

$$P_B = P(e\,|s_1)P(s_1) + P(e\,|s_2)P(s_2) \tag{3.37a}$$

or equivalently,

$$P_B = P(H_2|s_1)P(s_1) + P(H_1|s_2)P(s_2) \tag{3.37b}$$

That is, given that signal $s_1(t)$ was transmitted, an error results if hypothesis H_2 is chosen; or given that signal $s_2(t)$ was transmitted, an error results if hypothesis H_1 is chosen. For the case where the a priori probabilities are equal [that is, $P(s_1) = P(s_2) = \frac{1}{2}$],

$$P_B = \tfrac{1}{2}\,P(H_2|s_1) + \tfrac{1}{2}\,P(H_1|s_2) \tag{3.38}$$

and because of the symmetry of the probability density functions,

$$P_B = P(H_2|s_1) = P(H_1|s_2) \tag{3.39}$$

The probability of a bit error, P_B, is numerically equal to the area under the "tail" of either likelihood function, $p(z|s_1)$ or $p(z|s_2)$, falling on the "incorrect" side of the threshold. We can therefore compute P_B by integrating $p(z|s_1)$ between the limits $-\infty$ and γ_0, or by integrating $p(z|s_2)$ between the limits γ_0 and ∞:

$$P_B = \int_{\gamma_0 = (a_1 + a_2)/2}^{\infty} p(z\,|s_2)\,dz \tag{3.40}$$

Here, $\gamma_0 = (a_1 + a_2)/2$ is the optimum threshold from Equation (3.32). Replacing the likelihood $p(z|s_2)$ with its Gaussian equivalent from Equation (3.6), we have

$$P_B = \int_{\gamma_0=(a_1+a_2)/2}^{\infty} \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{z-a_2}{\sigma_0}\right)^2\right] dz \tag{3.41}$$

where σ_0^2 is the variance of the noise out of the correlator.

Let $u = (z - a_2)/\sigma_0$. Then $\sigma_0\, du = dz$ and

$$P_B = \int_{u=(a_1-a_2)/2\sigma_0}^{u=\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du = Q\left(\frac{a_1 - a_2}{2\sigma_0}\right) \tag{3.42}$$

where $Q(x)$, called the *complementary error function* or *co-error function,* is a commonly used symbol for the probability under the tail of the Gaussian pdf. It is defined as

$$Q(x) \approx \frac{1}{\sqrt{2\pi}} \int_x^{\infty} \exp\left(-\frac{u^2}{2}\right) du \tag{3.43}$$

Note that the co-error function is defined in several ways (see Appendix B); however, all definitions are equally useful for determining probability of error in Gaussian noise. $Q(x)$ cannot be evaluated in closed form. It is presented in tabular form in Table B.1. Good approximations to $Q(x)$ by simpler functions can be found in Reference [5]. One such approximation, valid for $x > 3$, is

$$Q(x) \approx \frac{1}{x\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \tag{3.44}$$

We have optimized (in the sense of minimizing P_B) the threshold level γ, but have not optimized the receiving filter in block 1 of Figure 3.1. We next consider optimizing this filter by maximizing the argument of $Q(x)$ in Equation (3.42).

3.2.2 The Matched Filter

A matched filter is a linear filter designed to provide the maximum signal-to-noise power ratio at its output for a given transmitted symbol waveform. Consider that a known signal $s(t)$ plus AWGN $n(t)$ is the input to a linear, time-invariant (receiving) filter followed by a sampler, as shown in Figure 3.1. At time $t = T$, the sampler output $z(T)$ consists of a signal component a_i and a noise component n_0. The variance of the output noise (average noise power) is denoted by σ_0^2, so that the ratio of the instantaneous signal power to average noise power, $(S/N)_T$, at time $t = T$, out of the sampler in step 1, is

$$\left(\frac{S}{N}\right)_T = \frac{a_i^2}{\sigma_0^2} \tag{3.45}$$

We wish to find the filter transfer function $H_0(f)$ that *maximizes* Equation (3.45). We can express the signal $a_i(t)$ at the filter output in terms of the filter transfer function $H(f)$ (before optimization) and the Fourier transform of the input signal, as

$$a_i(t) = \int_{-\infty}^{\infty} H(f)S(f)e^{j2\pi ft}\,df \tag{3.46}$$

where $S(f)$ is the Fourier transform of the input signal, $s(t)$. If the two-sided power spectral density of the input noise is $N_0/2$ watts/hertz, then, using Equations (1.19) and (1.53), we can express the output noise power as

$$\sigma_0^2 = \frac{N_0}{2}\int_{-\infty}^{\infty} |H(f)|^2\,df \tag{3.47}$$

We then combine Equations (3.45) to (3.47) to express $(S/N)_T$, as follows:

$$\left(\frac{S}{N}\right)_T = \frac{\left|\int_{-\infty}^{\infty} H(f)S(f)e^{j2\pi fT}\,df\right|^2}{N_0/2\int_{-\infty}^{\infty} |H(f)|^2\,df} \tag{3.48}$$

We next find that value of $H(f) = H_0(f)$ for which the maximum $(S/N)_T$ is achieved, by using *Schwarz's inequality.* One form of the inequality can be stated as

$$\left|\int_{-\infty}^{\infty} f_1(x)f_2(x)\,dx\right|^2 \leq \int_{-\infty}^{\infty} |f_1(x)|^2\,dx \int_{-\infty}^{\infty} |f_2(x)|^2\,dx \tag{3.49}$$

The equality holds if $f_1(x) = kf_2^*(x)$, where k is an arbitrary constant and * indicates complex conjugate. If we identify $H(f)$ with $f_1(x)$ and $S(f)\,e^{j2\pi fT}$ with $f_2(x)$, we can write

$$\left|\int_{-\infty}^{\infty} H(f)S(f)e^{j2\pi fT}df\right|^2 \leq \int_{-\infty}^{\infty} |H(f)|^2\,df \int_{-\infty}^{\infty} |S(f)|^2\,df \tag{3.50}$$

Substituting into Equation (3.48) yields

$$\left(\frac{S}{N}\right)_T \leq \frac{2}{N_0}\int_{-\infty}^{\infty} |S(f)|^2\,df \tag{3.51}$$

or

$$\max\left(\frac{S}{N}\right)_T = \frac{2E}{N_0} \tag{3.52}$$

where the energy E of the input signal $s(t)$ is

$$E = \int_{-\infty}^{\infty} |S(f)|^2\,df \tag{3.53}$$

Thus, the maximum output $(S/N)_T$ depends on the input *signal energy* and the power spectral density of the noise, *not on the particular shape* of the waveform that is used.

The equality in Equation (3.52) holds only if the optimum filter transfer function $H_0(f)$ is employed, such that

$$H(f) = H_0(f) = kS^*(f)e^{-j2\pi fT} \tag{3.54}$$

or

$$h(t) = \mathcal{F}^{-1}\{kS^*(f)e^{-j2\pi fT}\} \tag{3.55}$$

Since $s(t)$ is a real-valued signal, we can write, from Equations (A.29) and (A.31),

$$h(t) = \begin{cases} ks(T-t) & 0 \le t \le T \\ 0 & \text{elsewhere} \end{cases} \tag{3.56}$$

Thus, the impulse response of a filter that produces the maximum output signal-to-noise ratio is the mirror image of the message signal $s(t)$, *delayed* by the symbol time duration T. Note that the delay of T seconds makes Equation (3.56) *causal;* that is, the delay of T seconds makes $h(t)$ a function of positive time in the interval $0 \le t \le T$. Without the delay of T seconds, the response $s(-t)$ is unrealizable because it describes a response as a function of negative time.

3.2.3 Correlation Realization of the Matched Filter

Equation (3.56) and Figure 3.7a illustrate the matched filter's basic property: The impulse response of the filter is a delayed version of the mirror image (rotated on the $t = 0$ axis) of the signal waveform. Therefore, if the signal waveform is $s(t)$, its mirror image is $s(-t)$, and the mirror image delayed by T seconds is $s(T - t)$. The output $z(t)$ of a causal filter can be described in the time domain as the convolution of a received input waveform $r(t)$ with the impulse response of the filter (see Section A.5):

$$z(t) = r(t) * h(t) = \int_0^t r(\tau)h(t-\tau)\,d\tau \tag{3.57}$$

Substituting $h(t)$ of Equation (3.56) into $h(t-\tau)$ of Equation (3.57) and arbitrarily setting the constant k equal to unity, we get

$$\begin{aligned} z(t) &= \int_0^t r(\tau)s[T-(t-\tau)]\,d\tau \\ &= \int_0^t r(\tau)s(T-t+\tau)\,d\tau \end{aligned} \tag{3.58}$$

When $t = T$, we can write Equation (3.58) as

$$z(T) = \int_0^T r(\tau)s(\tau)\,d\tau \tag{3.59}$$

The operation of Equation (3.59), the product integration of the received signal $r(t)$ with a replica of the transmitted waveform $s(t)$ over one symbol interval is known as the *correlation* of $r(t)$ with $s(t)$. Consider that a received signal $r(t)$ is correlated with each prototype signal $s_i(t)$ $(i = 1, \ldots, M)$, using a bank of M correlators. The signal $s_i(t)$ whose product integration or correlation with $r(t)$ yields the maximum

difference could have been predicted by the factor-of-2 difference in the coefficient of E_b/N_0 in Equation (3.70) compared with (3.71). In Chapter 4, it is shown that, with MF detection, *bandpass* antipodal signaling (e.g., binary phase-shift keying) has the same P_B performance as *baseband* antipodal signaling (e.g., bipolar pulses). It is also shown that, with MF detection, *bandpass* orthogonal signaling (e.g., orthogonal frequency-shift keying) has the same P_B performance as *baseband* orthogonal signaling (e.g., unipolar pulses).

3.3 INTERSYMBOL INTERFERENCE

Figure 3.15a introduces the filtering aspects of a typical digital communication system. There are various filters (and reactive circuit elements such as inductors and capacitors) throughout the system—in the transmitter, in the receiver, and in the channel. At the transmitter, the information symbols, characterized as impulses or voltage levels, modulate pulses that are then filtered to comply with some bandwidth constraint. For baseband systems, the channel (a cable) has distributed reactances that distort the pulses. Some bandpass systems, such as wireless systems, are characterized by fading channels (see Chapter 15), that behave like undesirable filters manifesting signal distortion. When the receiving filter is configured to compensate for the distortion caused by *both* the transmitter and the channel, it is often referred to as an *equalizing filter* or a *receiving/equalizing filter*. Figure 3.15b illustrates a convenient model for the system, lumping all the filtering effects into one overall equivalent system transfer function

$$H(f) = H_t(f)\, H_c(f)\, H_r(f) \tag{3.77}$$

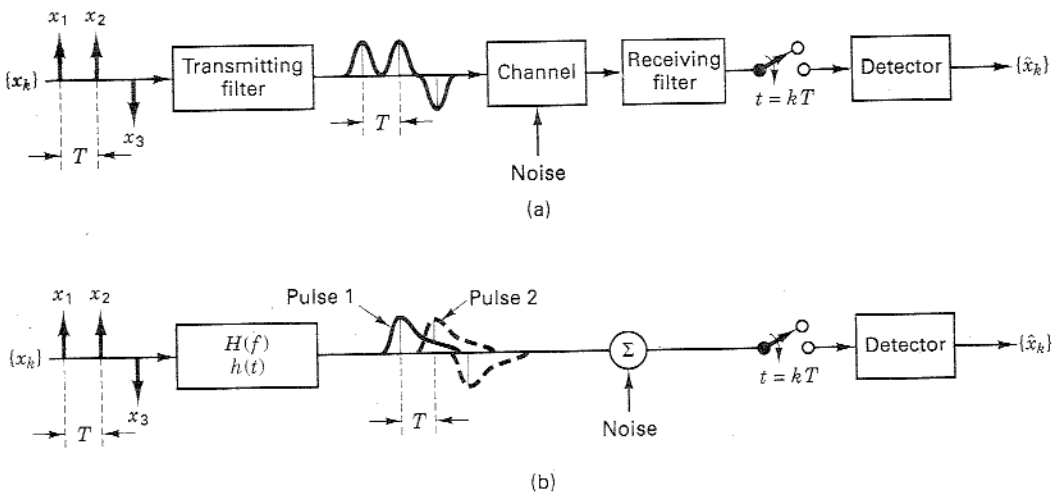

Figure 3.15 Intersymbol interference in the detection process. (a) Typical baseband digital system. (b) Equivalent model.

where $H_t(f)$ characterizes the transmitting filter, $H_c(f)$ the filtering within the channel, and $H_r(f)$ the receiving/equalizing filter. The characteristic $H(f)$, then, represents the composite system transfer function due to all the filtering at various locations throughout the transmitter/channel/receiver chain. In a binary system with a common PCM waveform, such as NRZ-L, the detector makes a symbol decision by comparing a sample of the received pulse to a threshold; for example, the detector in Figure 3.15 decides that a binary one was sent if the received pulse is positive, and that a binary zero was sent, if the received pulse is negative. Due to the effects of system filtering, the received pulses can overlap one another as shown in Figure 3.15b. The tail of a pulse can "smear" into adjacent symbol intervals, thereby interfering with the detection process and degrading the error performance; such interference is termed *intersymbol interference* (ISI). Even in the absence of noise, the effects of filtering and channel-induced distortion lead to ISI. Sometimes $H_c(f)$ is specified, and the problem remains to determine $H_t(f)$ and $H_r(f)$, such that the ISI is minimized at the output of $H_r(f)$.

Nyquist [6] investigated the problem of specifying a received pulse shape so that no ISI occurs at the detector. He showed that the theoretical minimum system bandwidth needed in order to detect R_s symbols/s, without ISI, is $R_s/2$ hertz. This occurs when the system transfer function $H(f)$ is made rectangular, as shown in Figure 3.16a. For baseband systems, when $H(f)$ is such a filter with single-sided bandwidth $1/2T$ (the *ideal Nyquist filter*), its impulse response, the inverse Fourier transform of $H(f)$ (from Table A.1) is of the form $h(t) = \text{sinc}\,(t/T)$, shown in Figure 3.16b. This sinc (t/T)-shaped pulse is called the *ideal Nyquist pulse;* its multiple lobes comprise a mainlobe and sidelobes called pre- and post-mainlobe *tails* that are infinitely long. Nyquist established that if each pulse of a received sequence is of the form sinc (t/T), the pulses can be detected without ISI. Figure 3.16b illustrates how ISI is avoided. There are two successive pulses, $h(t)$ and $h(t - T)$. Even though $h(t)$ has long tails, the figure shows a tail passing through zero amplitude at the instant $(t = T)$ when $h\ (t - T)$ is to be sampled, and likewise all tails pass through zero amplitude when any other pulse of the sequence $h(t - kT)$, $k = \pm 1$, $\pm 2, \ldots$ is to be sampled. Therefore, assuming that the sample timing is perfect, there will be no ISI degradation introduced. For baseband systems, the bandwidth

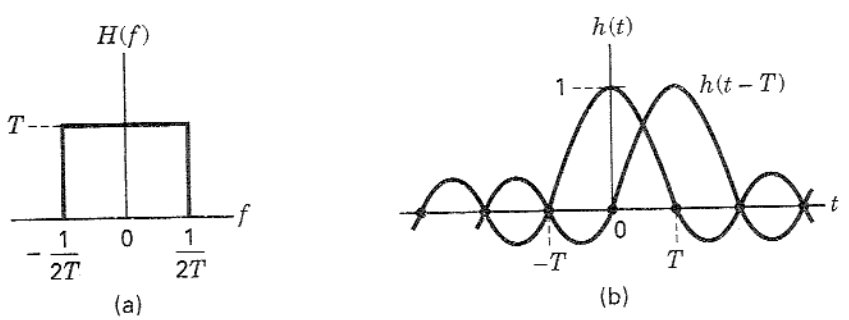

Figure 3.16 Nyquist channels for zero ISI. (a) Rectangular system transfer function $H(f)$. (b) Received pulse shape $h(t) = \text{sinc}\,(t/T)$.

required to detect $1/T$ such pulses (symbols) per second is equal to $1/2T$; in other words, a system with bandwidth $W = 1/2T = R_s/2$ hertz can support a maximum transmission rate of $2W = 1/T = R_s$ symbols/s (*Nyquist bandwidth constraint*) without ISI. Thus, for ideal Nyquist filtering (and zero ISI), the maximum possible symbol transmission rate per hertz, called the *symbol-rate packing,* is 2 symbols/s/Hz. It should be clear from the rectangular-shaped transfer function of the ideal Nyquist filter and the infinite length of its corresponding pulse, that such ideal filters are not realizable; they can only be approximately realized.

The names "Nyquist filter" and "Nyquist pulse" are often used to describe the general class of filtering and pulse-shaping that satisfy zero ISI at the sampling points. A Nyquist filter is one whose frequency transfer function can be represented by a rectangular function convolved with any real even-symmetric frequency function. A Nyquist pulse is one whose shape can be represented by a sinc (t/T) function multiplied by another time function. Hence, there are a countless number of Nyquist filters and corresponding pulse shapes. Amongst the class of Nyquist filters, the most popular ones are the raised cosine and the root-raised cosine, treated below.

A fundamental parameter for communication systems is *bandwidth efficiency, R/W,* whose units are bits/s/Hz. As the units imply, R/W represents a measure of data throughput per hertz of bandwidth and thus measures how efficiently any signaling technique utilizes the bandwidth resource. Since the Nyquist bandwidth constraint dictates that the theoretical maximum symbol-rate packing without ISI is 2 symbols/s/Hz, one might ask what it says about the maximum number of bits/s/Hz. It says nothing about bits, directly; the constraint deals only with pulses or symbols, and the ability to detect their amplitude values without distortion from other pulses. To find R/W for any signaling scheme, one must know how many bits each symbol represents, which is a separate issue. Consider an M-ary PAM signaling set. Each symbol (comprising k bits) is represented by one of M-pusle amplitudes. For $k = 6$ bits per symbol, the symbol set size is $M = 2^k = 64$ amplitudes. Thus with 64-ary PAM, the theoretical maximum bandwidth efficiency that is possible without ISI is 12 bits/s/Hz. (Bandwidth efficiency is treated in greater detail in Chapter 9.)

3.3.1 Pulse Shaping to Reduce ISI

3.3.1.1 Goals and Trade-offs

The more compact we make the signaling spectrum, the higher is the allowable data rate or the greater is the number of users that can simultaneously be served. This has important implications to communication service providers, since greater utilization of the available bandwidth translates into greater revenue. For most communication systems (with the exception of spread-spectrum systems, covered in Chapter 12), our goal is to reduce the required system bandwidth as much as possible. Nyquist has provided us with a basic limitation to such bandwidth reduction. What would happen if we tried to force a system to operate at smaller bandwidths than the constraint dictates? The pulses would become spread in time,

which would degrade the system's error performance due to increased ISI. A prudent goal is to compress the bandwidth of the data impulses to some reasonably small bandwidth greater than the Nyquist minimum. This is accomplished by pulse-shaping with a Nyquist filter. If the band edge of the filter is steep, approaching the rectangle in Figure 3.16a, then the signaling spectrum can be made most compact. However, such a filter has an impulse response duration approaching infinity, as indicated in Figure 3.16b. Each pulse extends into every pulse in the entire sequence. Long time responses exhibit large-amplitude tails nearest the main lobe of each pulse. Such tails are undesirable because, as shown in Figure 3.16b, they contribute zero ISI *only* when the sampling is performed *at exactly* the correct sampling time; when the tails are large, small timing errors will result in ISI. Therefore, although a compact spectrum provides optimum bandwidth utilization, it is very susceptible to ISI degradation induced by timing errors.

3.3.1.2 The Raised-Cosine Filter

Earlier, it was stated that the receiving filter is often referred to as an *equalizing filter,* when it is configured to compensate for the distortion caused by both the transmitter and the channel. In other words, the configuration of this filter is chosen so as to optimize the composite system frequency transfer function $H(f)$, shown in Equation (3.77). One frequently used $H(f)$ transfer function belonging to the Nyquist class (zero ISI at the sampling times) is called the *raised-cosine filter.* It can be expressed as

$$H(f) = \begin{cases} 1 & \text{for } |f| < 2W_0 - W \\ \cos^2\left(\dfrac{\pi}{4}\dfrac{|f| + W - 2W_0}{W - W_0}\right) & \text{for } 2W_0 - W < |f| < W \\ 0 & \text{for } |f| > W \end{cases} \quad (3.78)$$

where W is the absolute bandwidth and $W_0 = 1/2T$ represents the minimum Nyquist bandwidth for the rectangular spectrum and the –6-dB bandwidth (or half-amplitude point) for the raised-cosine spectrum. The difference $W - W_0$ is termed the "excess bandwidth," which means additional bandwidth beyond the Nyquist minimum (i.e., for the rectangular spectrum, W is equal to W_0). The *roll-off factor* is defined to be $r = (W - W_0)/W_0$, where $0 \le r \le 1$. It represents the excess bandwidth divided by the filter –6-dB bandwidth (i.e., the fractional excess bandwidth). For a given W_0, the roll-off r specifies the required excess bandwidth as a fraction of W_0 and characterizes the steepness of the filter roll off. The raised-cosine characteristic is illustrated in Figure 3.17a for roll-off values of $r = 0$, $r = 0.5$, and $r = 1$. The $r = 0$ roll-off is the Nyquist minimum-bandwidth case. Note that when $r = 1$, the required excess bandwidth is 100%, and the tails are quite small. A system with such an overall spectral characteristic can provide a symbol rate of R_s symbols/s using a bandwidth of R_s hertz (twice the Nyquist minimum bandwidth), thus yielding a symbol-rate packing of 1 symbol/s/Hz. The corresponding impulse response for the $H(f)$ of Equation (3.78) is

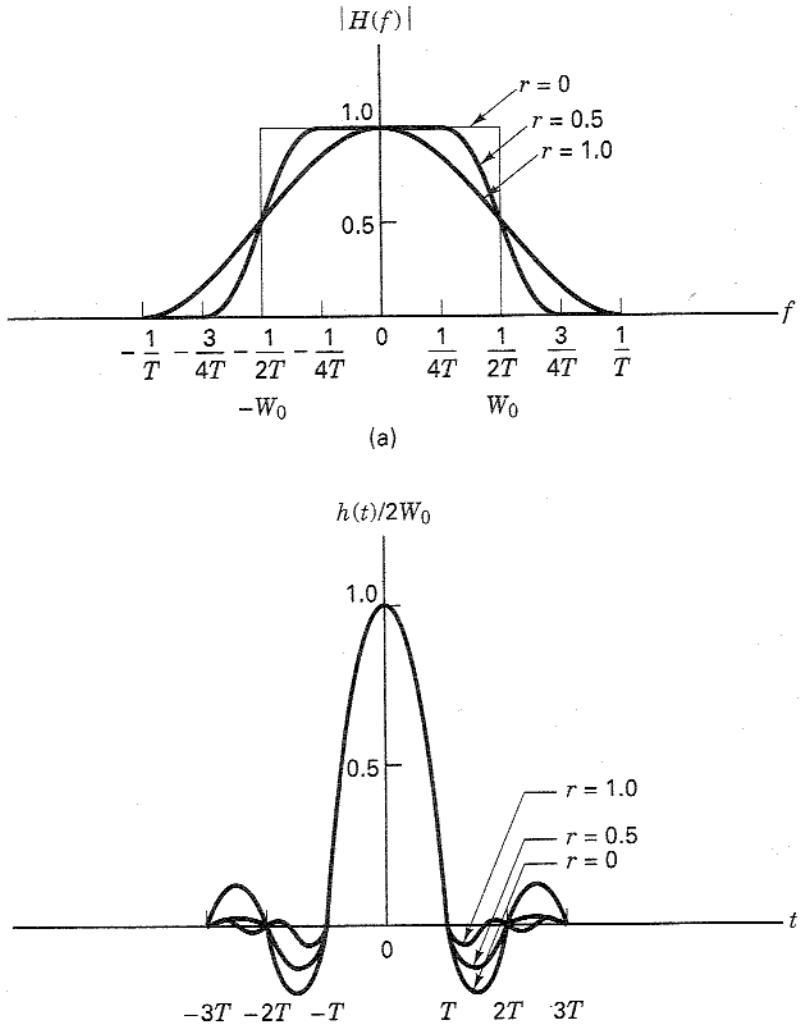

Figure 3.17 Raised-cosine filter characteristics. (a) System transfer function. (b) System impulse response.

$$h(t) = 2W_0(\text{sinc}\, 2W_0 t)\frac{\cos\,[2\pi(W - W_0)t]}{1 - [4(W - W_0)t]^2} \tag{3.79}$$

and is plotted in Figure 3.17b for $r = 0$, $r = 0.5$, and $r = 1$. The tails have zero value at each pulse-sampling time, regardless of the roll-off value.

We can only approximately implement a filter described by Equation (3.78) and a pulse shape described by Equation (3.79), since, strictly speaking, the raised-cosine spectrum is not physically realizable (for the same reason that the ideal Nyquist filter is not realizable). A realizable filter must have an impulse response of finite duration and exhibit a zero output prior to the pulse turn-on time (see Sec-

tion 1.7.2), which is not the case for the family of raised-cosine characteristics. These unrealizable filters are *noncausal* (the filter impulse response has infinite duration, and the filtered pulse begins at time $t = -\infty$). A pulse-shaping filter should satisfy two requirements. It should provide the desired roll-off, and it should be realizable (the impulse response needs to be truncated to a finite length).

Starting with the Nyquist bandwidth constraint that the minimum required system bandwidth W for a symbol rate of R_s symbols/s without ISI is $R_s/2$ hertz, a more general relationship between required bandwidth and symbol transmission rate involves the filter roll-off factor r and can be stated as

$$W = \frac{1}{2}(1 + r)R_s \tag{3.80}$$

Thus, with $r = 0$, Equation (3.80) describes the minimum required bandwidth for ideal Nyquist filtering. For $r > 0$, there is a bandwidth expansion beyond the Nyquist minimum; thus, for this case, R_s is now less than twice the bandwidth. If the demodulator outputs one sample per symbol, then the Nyquist sampling theorem has been violated, since we are left with too few samples to reconstruct the analog waveform unambiguously (aliasing is present). However, for digital communication systems, we are not interested in reconstructing the analog waveform. Since the family of raised-cosine filters is characterized by zero ISI at the times that the symbols are sampled, we can still achieve unambiguous detection.

Bandpass-modulated signals (see Chapter 4), such as amplitude shift keying (ASK) and phase-shift keying (PSK), require twice the transmission bandwidth of the equivalent baseband signals. (See Section 1.7.1.) Such frequency-translated signals, occupying twice their baseband bandwidth, are often called double-sideband (DSB) signals. Therefore, for ASK- and PSK-modulated signals, the relationship between the required DSB bandwidth W_{DSB} and the symbol transmission rate R_s is

$$W_{DSB} = (1 + r)R_s \tag{3.81}$$

Recall that the raised-cosine frequency transfer function describes the composite $H(f)$ that is the "full round trip" from the inception of the message (as an impulse) at the transmitter, through the channel, and through the receiving filter. The filtering at the receiver is the compensating portion of the overall transfer function to help bring about zero ISI with an overall transfer function, such as the raised cosine. Often this is accomplished by choosing (matching) the receiving filter and the transmitting filter so that each has a transfer function known as a root-raised cosine (square root of the raised cosine). Neglecting any channel-induced ISI, the product of these root-raised-cosine functions yields the composite raised-cosine system transfer function. Whenever a separate equalizing filter is introduced to mitigate the effects of channel-induced ISI, the receiving and equalizing filters together should be configured to compensate for the distortion caused by both the transmitter and the channel so as to yield an overall system transfer function characterized by zero ISI.

Let's review the trade-off that faces us in specifying pulse-shaping filters. The larger the filter roll-off, the shorter will be the pulse tails (which implies smaller tail

amplitudes). Small tails exhibit less sensitivity to timing errors and thus make for small degradation due to ISI. Notice in Figure 3.17b, for $r = 1$, that timing errors can still result in some ISI degradation. However, the problem is not as serious as it is for the case in which $r = 0$, because the tails of the $h(t)$ waveform are of much smaller amplitude for $r = 1$ than they are for $r = 0$. The cost is more excess bandwidth. On the other hand, the smaller the filter roll-off, the smaller will be the excess bandwidth, thereby allowing us to increase the signaling rate or the number of users that can simultaneously use the system. The cost is longer pulse tails, larger pulse amplitudes, and thus, greater sensitivity to timing errors.

3.3.2 Two Types of Error-Performance Degradation

The effects of error-performance degradation in digital communications can be partitioned into two categories. The first is due to a decrease in received signal power or an increase in noise or interference power, giving rise to a loss in signal-to-noise ratio or E_b/N_0. The second is due to signal distortion, such as might be caused by intersymbol interference (ISI). Let us demonstrate how different are the effects of these two degradation types.

Suppose that we require a communication system with a bit-error probability P_B versus E_b/N_0 characteristic corresponding to the solid-line curve plotted in Figure 3.18a. Suppose that after the system is configured and measurements are taken, we find, to our disappointment, that the performance does not follow the theoretical curve, but in fact follows the dashed line plot. A loss in E_b/N_0 has come about because of some signal losses or an increased level of noise or interference. For a

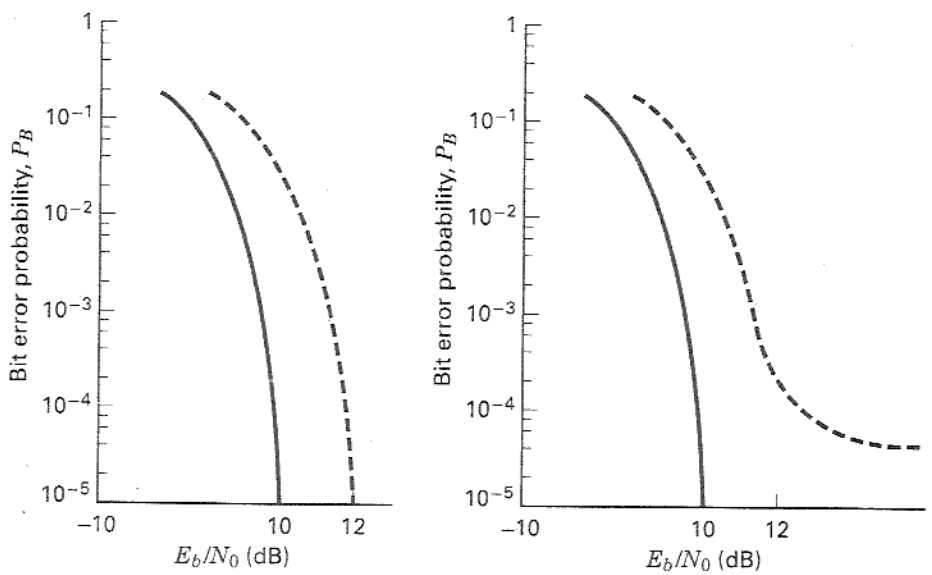

Figure 3.18 (a) Loss in E_b/N_0. (b) Irreducible P_B caused by distortion.

desired bit-error probability of 10^{-5}, the theoretical required E_b/N_0 is 10 dB. Since our system performance falls short of our goal, we can see from the dashed-line curve that, for the same bit-error probability of 10^{-5}, the required E_b/N_0 is now 12 dB. If there were no way to remedy this problem, how much more E_b/N_0 would have to be provided in order to meet the required bit-error probability? The answer is 2 dB, of course. It might be a serious problem—especially if the system is power-limited, and it is difficult to come up with the additional 2 dB. But that loss in E_b/N_0 is not *so* terrible when compared with the possible effects of degradation caused by a distortion mechanism.

In Figure 3.18b, again imagine that we do not meet the desired performance of the solid-line curve. But instead of suffering a simple loss in signal-to-noise ratio, there is a degradation effect brought about by ISI (plotted with the dashed line). If there were no way to remedy this problem, how much more E_b/N_0 would be required in order to meet the desired bit-error probability? It would require an infinite amount—or, in other words, there is no amount of E_b/N_0 that will ameliorate this problem. More E_b/N_0 cannot help when the curve manifests such an irreducible P_B (assuming that the bottoming-out point is located above the system's required P_B). Undoubtedly, every P_B-versus-E_b/N_0 curve bottoms out somewhere, but if the bottoming-out point is well below the region of interest, it will be of no consequence.

More E_b/N_0 may not help the ISI problem (it won't help at all if the P_B curve has reached an irreducible level). This can be inferred by looking at the overlapped pulses in Figure 3.15b; if we increase the E_b/N_0, the ratio of that overlap does not change. The pulses are subject to the same distortion. What, then, is the usual cure for the degradation effects of ISI? The cure is found in a technique called equalization. (See Section 3.4.) Since the distortion effects of ISI are caused by filtering in the transmitter and the channel, equalization can be thought of as the process that reverses such nonoptimum filtering effects.

Example 3.3 Bandwidth Requirements

(a) Find the minimum required bandwidth for the baseband transmission of a four-level PAM pulse sequence having a data rate of $R = 2400$ bits/s if the system transfer characteristic consists of a raised-cosine spectrum with 100% excess bandwidth ($r = 1$).

(b) The same 4-ary PAM sequence is modulated onto a carrier wave, so that the baseband spectrum is shifted and centered at frequency f_0. Find the minimum required DSB bandwidth for transmitting the modulated PAM sequence. Assume that the system transfer characteristic is the same as in part (a).

Solution

(a) $M = 2^k$; since $M = 4$ levels, $k = 2$.

$$\text{Symbol or pulse rate } R_s = \frac{R}{k} = \frac{2400}{2} = 1200 \text{ symbols/s;}$$

$$\text{Minimum bandwidth } W = \tfrac{1}{2}(1 + r)R_s = \tfrac{1}{2}(2)(1200) = 1200 \text{ Hz.}$$

Figure 3.19a illustrates the baseband PAM received pulse in the time domain—an approximation to the $h(t)$ in Equation (3.79). Figure 3.19b illustrates the Fourier

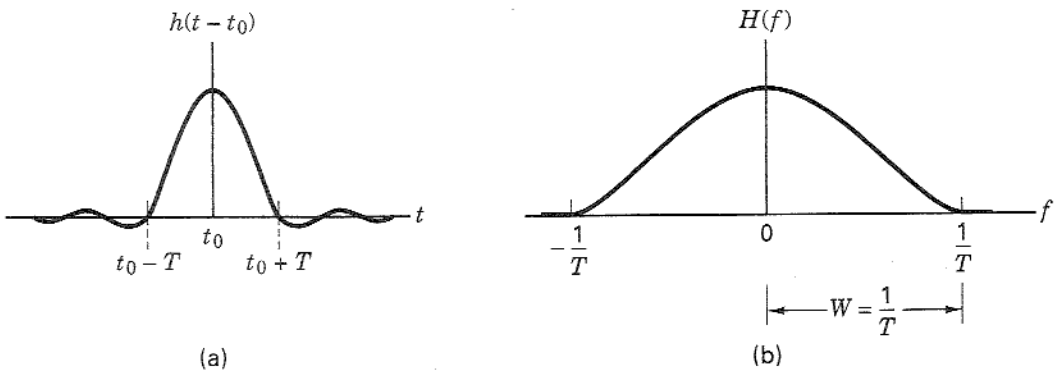

Figure 3.19 (a) Shaped pulse. (b) Baseband raised cosine spectrum.

transform of $h(t)$—the raised cosine spectrum. Notice that the required bandwidth, W, starts at zero frequency and extends to $f = 1/T$; it is twice the size of the Nyquist theoretical minimum bandwidth.

(b) As in part (a),

$$R_s = 1200 \text{ symbols/s};$$

$$W_{\text{DSB}} = (1 + r)R_s = 2(1200) = 2400 \text{ Hz}.$$

Figure 3.20a illustrates the modulated PAM received pulse. This waveform can be viewed as the product of a high-frequency sinusoidal carrier wave and a waveform with the pulse shape of Figure 3.19a. The single-sided spectral plot in Figure 3.20b illustrates that the modulated bandwidth is

$$W_{\text{DSB}} = \left(f_0 + \frac{1}{T}\right) - \left(f_0 - \frac{1}{T}\right) = \frac{2}{T}.$$

When the spectrum of Figure 3.19b is shifted up in frequency, the negative and positive halves of the baseband spectrum are shifted up in frequency, thereby dou-

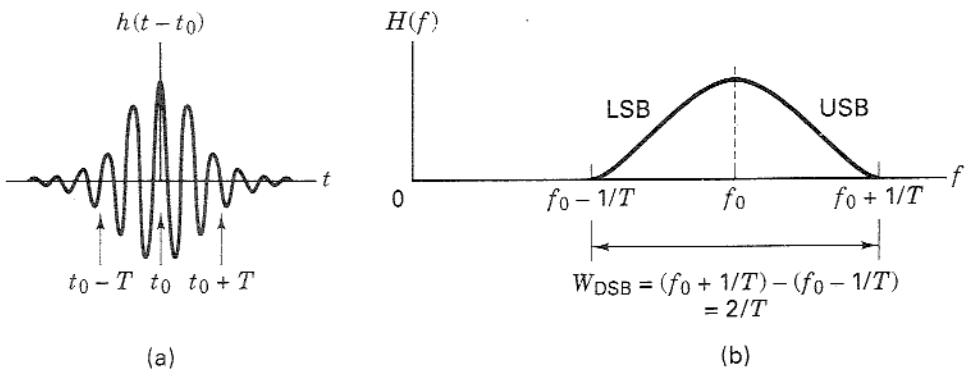

Figure 3.20 (a) Modulated shaped pulse. (b) DSB-modulated raised cosine spectrum.

bling the required transmission bandwidth. As the name implies, the DSB signal has two sidebands: the upper sideband (USB), derived from the baseband positive half, and the lower sideband (LSB), derived from the baseband negative half.

Example 3.4 Digital Telephone Circuits

Compare the system bandwidth requirements for a terrestrial 3-kHz analog telephone voice channel with that of a digital one. For the digital channel, the voice is formatted as a PCM bit stream, where the sampling rate for the analog-to-digital (A/D) conversion is 8000 samples/s and each voice sample is quantized to one of 256 levels. The bit stream is then transmitted using a PCM waveform and received with zero ISI.

Solution

The result of the sampling and quanitization process yields PCM words such that each word (representing one sample) has one of $L = 256$ different levels. If each sample were sent as a 256-ary PAM pulse (symbol), then from Equation (3.80) we can write that the required system bandwidth (without ISI) for sending R_s symbols/s would be

$$W \geq \frac{R_s}{2} \text{ hertz}$$

where the equality sign holds true only for ideal Nyquist filtering. Since the digital telephone system uses PCM (binary) waveforms, each PCM word is converted to $\ell = \log_2 L = \log_2 256 = 8$ bits. Therefore, the system bandwidth required to transmit voice using PCM is

$$W_{\text{PCM}} \geq (\log_2 L)\frac{R_s}{2} \text{ hertz}$$

$$\geq \frac{1}{2}(8 \text{ bits/symbol})\,(8000 \text{ symbols/s}) = 32 \text{ kHz}$$

The 3-kHz analog voice channel will generally require approximately 4-kHz of bandwidth, including some bandwidth separation between channels, called *guard bands.* Therefore, the PCM format, using 8-bit quantization and binary signaling with a PCM waveform, requires at least eight times the bandwidth required for the analog channel.

3.3.3 Demodulation/Detection of Shaped Pulses

3.3.3.1 Matched Filters versus Conventional Filters

Conventional filters screen out unwanted spectral components of a received signal while maintaining some measure of fidelity for signals occupying a selected span of the spectrum, called the *pass-band.* These filters are generally designed to provide approximately uniform gain, a linear phase-versus-frequency characteristic over the pass-band, and a specified minimum attenuation over the remaining spectrum, called the *stop-band(s).* A matched filter has a different "design priority," namely that of maximizing the SNR of a known signal in the presence of AWGN. Conventional filters are applied to random signals defined only by their bandwidth, while matched filters are applied to *known signals* with random parameters (such as amplitude and arrival time). The matched filter can be considered to be a *template* that is matched to the known shape of the signal being processed. A conven-

tional filter tries to preserve the temporal or spectral structure of the signal of interest. On the other hand, a matched filter significantly modifies the temporal structure by gathering the signal energy matched to its template, and, at the end of each symbol time, presenting the result as a peak amplitude. In general, a digital communications receiver processes received signals with both kinds of filters. The task of the conventional filter is to isolate and extract a high-fidelity estimate of the signal for presentation to the matched filter. The matched filter gathers the received signal energy, and when its output is sampled (at $t = T$), a voltage proportional to that energy is produced for subsequent detection and post-detection processing.

3.3.3.2 Nyquist Pulse and Square-Root Nyquist Pulse

Consider a sequence of data impulses at a transmitter input compared with the resulting sequence of pulses out of a raised-cosine matched filter (before sampling). In Figure 3.21, transmitted data is represented by impulse waveforms that occur at times $\tau_0, \tau_1, \ldots$ Filtering spreads the input waveforms, and thus delays them in time. We use the notation, $t_0, t_1, \ldots,$ to denote received time. The impulse event that was transmitted at time τ_0 arrives at the receiver at time t_0 corresponding to the start of the output pulse event. The premainlobe tail of a demodulated pulse is referred to as its *precursor.* For a real system with a fixed system-time reference, causality dictates that $t_0 \geq \tau_0$, and the time difference between τ_0 and t_0 represents any propagation delay in the system. In this example, the time duration from the start of a demodulated pulse precursor until the appearance of its mainlobe or peak amplitude is $3T$ (three pulse-time durations). Each output pulse in the sequence is superimposed with other pulses; each pulse has an effect on the main lobes of three earlier and three later pulses. When a pulse is filtered (shaped) so that it occupies more than one symbol time, we define the pulse *support time* as the total number of symbol intervals over which the pulse persists. In Figure 3.21, the pulse support time consists of 6-symbol intervals (7 data points with 6 intervals between them).

The impulse response of a root-raised cosine filter, called the *square-root Nyquist pulse,* is shown in Figure 3.22a (normalized to a peak value of unity, with a filter rolloff of $r = 0.5$). The impulse response of the raised-cosine filter, called the

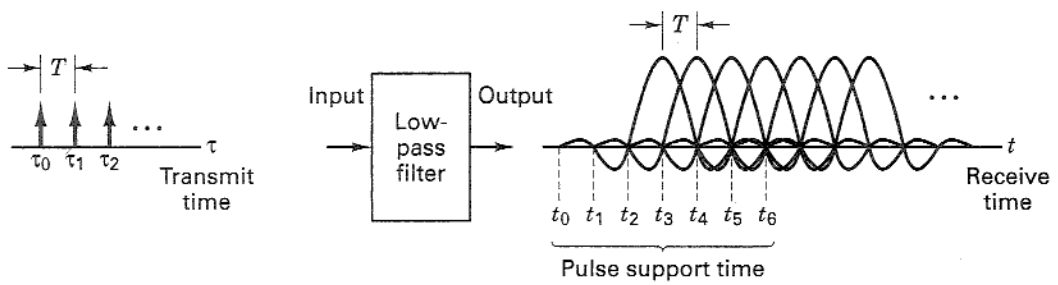

Figure 3.21 Filtered impulse sequence: output versus input.

Nyquist pulse, is shown in Figure 3.22b (with the same normalization and filter rolloff). Inspecting these two pulse shapes, we see that they have a similar appearance, but the square-root Nyquist pulse makes slightly faster transitions, thus its spectrum (root-raised cosine) does not decay as rapidly as the spectrum (raised cosine) of the Nyquist pulse. Another subtle but important difference is that the square-root Nyquist pulse *does not* exhibit zero ISI (you can verify that the pulse tails in Figure 3.22a do not go through zero amplitude at the symbol times). However, if a root-raised cosine filter is used at both the transmitter and the receiver, the product of these transfer functions being a raised cosine, will give rise to an output having zero ISI.

It is interesting to see how the square-root Nyquist pulses appear at the output of a transmitter and how they appear after demodulation with a root-raised cosine MF. Figure 3.23a illustrates an example of sending a sequence of message symbols {+1 +1 −1 +3 +1 +3} from a 4-ary set, where the members of the alphabet set are: {±1, ±3}. Consider that the pulse modulation is 4-ary PAM, and that the pulses have been shaped with a root-raised cosine filter, having a roll-off value of 0.5. The analog waveform in this figure represents the transmitter output. Since the output waveform from any filter is delayed in time, then in Figure 3.23a, the input message symbols (shown as approximate impulses) have been delayed the same amount as the output waveform in order to align the message sequence with its corresponding filtered waveform (the square-root Nyquist shaped-pulse sequence). This is just a visual convenience so that the reader can compare the filter input with its output. It is, of course, only the output analog waveform that is transmitted (or modulated) onto a carrier wave.

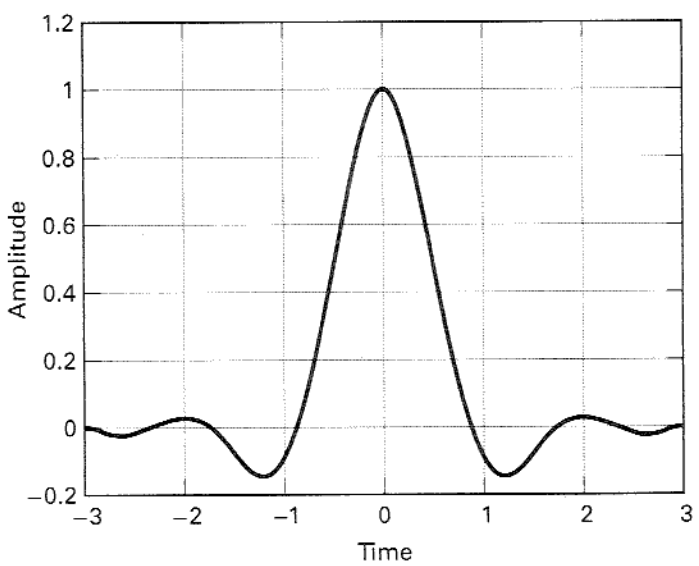

Figure 3.22a Square-root Nyquist pulse.

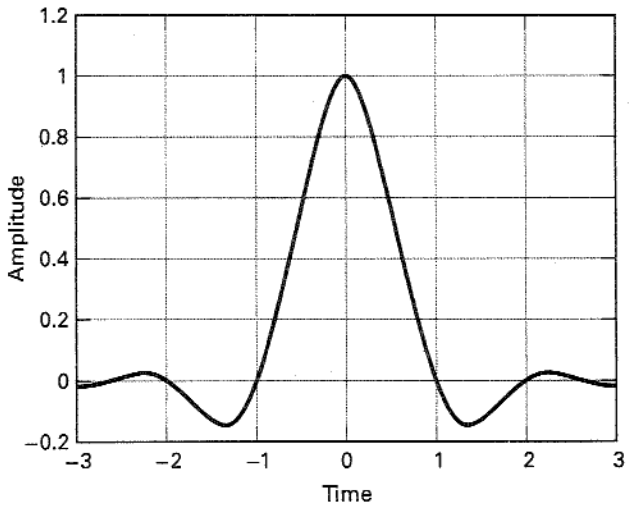

Figure 3.22b Nyquist pulse.

Figure 3.23b shows the same delayed message samples together with the output waveform from the root-raised cosine MF, yielding a raised-cosine transfer function for the overall system. Let us describe a simple test to determine if the filtered output (assuming no noise) contains ISI. It is only necessary to sample the filtered waveform at the times corresponding to the original input samples; if

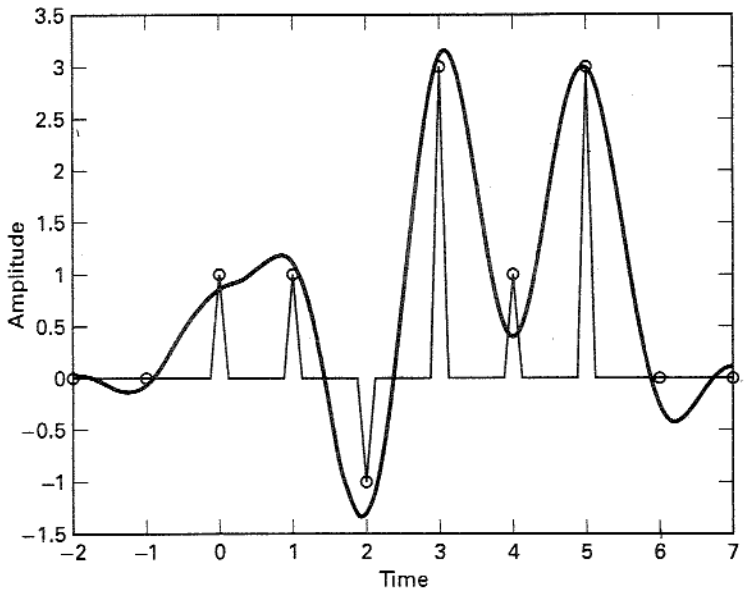

Figure 3.23a Square-root Nyquist-shaped *M*-ary waveform and delayed-input sample values.

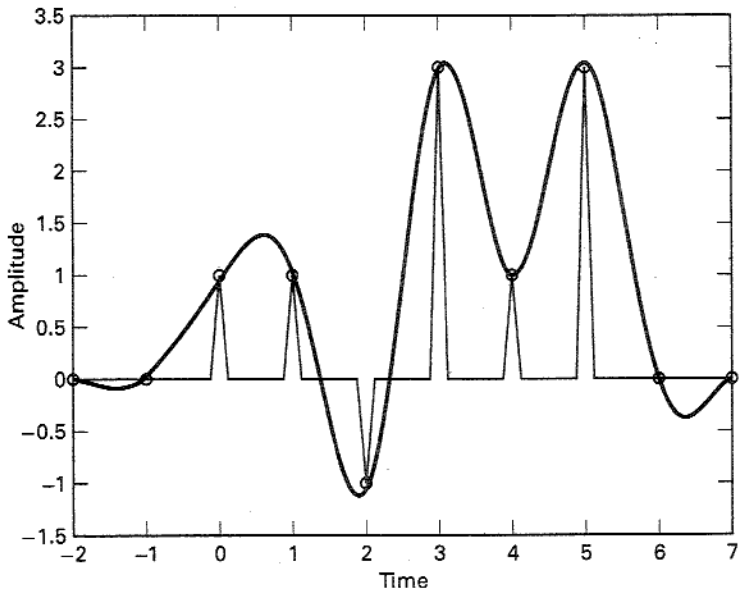

Figure 3.23b Output of raised-cosine matched filter and delayed-input sample values.

the resulting sample values are unchanged from those of the original message, then the filter output has zero ISI (at the sample times). When Figures 3.23a and 3.23b are compared with regard to ISI, it should be apparent that sampling the square-root Nyquist waveform of Figure 3.23a (transmitter output) *will not* yield the exact original samples; however, sampling the Nyquist waveform in Figure 3.23b (MF output) *will* yield the exact original samples. This supports the statement that a Nyquist filter yields zero ISI at the sample points, while any other filter does not do so.

3.4 EQUALIZATION

3.4.1 Channel Characterization

Many communication channels (e.g., telephone, wireless) can be characterized as band-limited linear filters with an impulse response $h_c(t)$ and a frequency response

$$H_c(f) = |H_c(f)|e^{j\theta_c(f)} \tag{3.82}$$

where $h_c(t)$ and $H_c(f)$ are Fourier transform pairs, $|H_c(f)|$ is the channel's amplitude response, and $\theta_c(f)$ is the channel's phase response. In order to achieve ideal (nondistorting) transmission characteristics over a channel, it was shown in Section 1.6.3, that within a signal's bandwidth W, $|H_c(f)|$ *must be constant.* Also, $\theta_c(f)$ *must be a linear function of frequency,* which is tantamount to saying that the delay must be constant for all spectral components of the signal. If $|H_c(f)|$ is not constant

within W, then the effect is amplitude distortion. If $\theta_c(f)$ is not a linear function of frequency within W, then the effect is phase distortion. For many channels that exhibit distortion of this type, such as fading channels, amplitude and phase distortion typically occur together. For a transmitted sequence of pulses, such distortion manifests itself as a signal dispersion or "smearing" so that any one pulse in the received demodulated sequence is not well defined. The overlap or smearing, known as *intersymbol interference* (ISI), described in Section 3.3, arises in most modulation systems; it is one of the major obstacles to reliable high-speed data transmission over bandlimited channels. In the broad sense, the name "equalization" refers to any signal processing or filtering technique that is designed to eliminate or reduce ISI.

In Figure 2.1, equalization is partitioned into two broad categories. The first category, *maximum-likelihood sequence estimation* (MLSE), entails making measurements of $h_c(t)$ and then providing a means for adjusting the receiver to the transmission environment. The goal of such adjustments is to enable the detector to make good estimates from the demodulated distorted pulse sequence. With an MLSE receiver, the distorted samples are not reshaped or directly compensated in any way; instead, the mitigating technique for the MLSE receiver is to adjust itself in such a way that it can better deal with the distorted samples. (An example of this method, known as Viterbi equalization, is treated in Section 15.7.1.) The second category, *equalization with filters,* uses filters to compensate the distorted pulses. In this second category, the detector is presented with a sequence of demodulated samples that the equalizer has modified or "cleaned up" from the effects of ISI. Equalizing with filters, the more popular approach and the one described in this section, lends itself to further partitioning. The filters can be described as to whether they are linear devices that contain only feedforward elements *transversal equalizers*), or whether they are nonlinear devices that contain both feedforward and feedback elements (*decision feedback equalizers*). They can be grouped according to the automatic nature of their operation, which may be either *preset* or *adaptive.* They also can be grouped according to the filter's resolution or update rate. Are predetection samples provided only on symbol boundaries, that is, one sample per symbol? If so, the condition is known as *symbol spaced.* Are multiple samples provided for each symbol? If so, this condition is known as *fractionally spaced.*

We now modify Equation (3.77) by letting the receiving/equalizing filter be replaced by a separate receiving filter and equalizing filter, defined by frequency transfer functions $H_r(f)$ and $H_e(f)$, respectively. Also, let the overall system transfer function $H(f)$ be a raised-cosine filter, designated $H_{RC}(f)$. Thus, we write

$$H_{RC}(f) = H_t(f)\, H_c(f)\, H_r(f)\, H_e(f) \tag{3.83}$$

In practical systems, the channel's frequency transfer function $H_c(f)$ and its impulse response $h_c(t)$ are not known with sufficient precision to allow for a receiver design to yield zero ISI for all time. Usually, the transmit and receive filters are chosen to be matched so that

$$H_{RC}(f) = H_t(f)\, H_r(f) \tag{3.84}$$

In this way, $H_t(f)$ and $H_r(f)$ each have frequency transfer functions that are the square root of the raised cosine (root-raised cosine). Then, the equalizer transfer function needed to compensate for channel distortion is simply the inverse of the channel transfer function:

$$H_e(f) = \frac{1}{H_c(f)} = \frac{1}{|H_c(f)|} e^{-j\theta_c(f)} \tag{3.85}$$

Sometimes a system frequency transfer function manifesting ISI at the sampling points is purposely chosen (e.g., a Gaussian filter transfer function). The motivation for such a transfer function is to improve bandwidth efficiency, compared with using a raised-cosine filter. When such a design choice is made, the role of the equalizing filter is not only to compensate for the channel-induced ISI, but also to compensate for the ISI brought about by the transmitter and receiver filters [7].

3.4.2 Eye Pattern

An eye pattern is the display that results from measuring a system's response to baseband signals in a prescribed way. On the vertical plates of an oscilloscope we connect the receiver's response to a random pulse sequence. On the horizontal plates we connect a sawtooth wave at the signaling frequency. In other words, the horizontal time base of the oscilloscope is set equal to the symbol (pulse) duration. This setup superimposes the waveform in each signaling interval into a family of traces in a single interval $(0, T)$. Figure 3.24 illustrates the eye pattern that results for binary antipodal (bipolar pulse) signaling. Because the symbols stem from a random source, they are sometimes positive and sometimes negative, and the persistence of the cathode ray tube display allows us to see the resulting pattern shaped as an eye. The width of the opening indicates the time over which sampling for detection might be performed. Of course, the optimum sampling time corresponds to the maximum eye opening, yielding the greatest protection against noise.

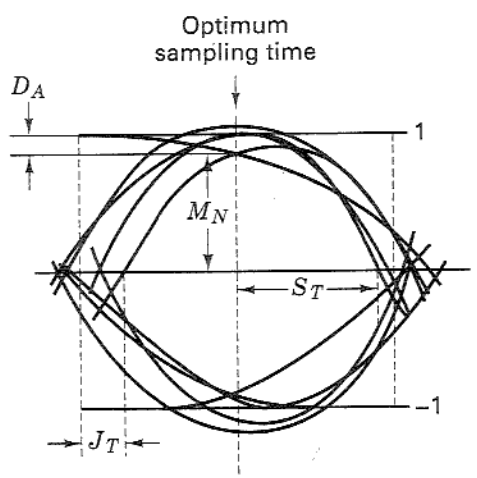

Figure 3.24 Eye Pattern.

If there were no filtering in the system—that is, if the bandwidth corresponding to the transmission of these data pulses were infinite—then the system response would yield ideal rectangular pulse shapes. In that case, the pattern would look like a box rather than an eye. In Figure 3.24, the range of amplitude differences labelled D_A is a measure of distortion caused by ISI, and the range of time differences of the zero crossings labelled J_T is a measure of the timing jitter. Measures of noise margin M_N and sensitivity-to-timing error S_T are also shown in the figure. In general, the most frequent use of the eye pattern is for qualitatively assessing the extent of the ISI. As the eye closes, ISI is increasing; as the eye opens, ISI is decreasing.

3.4.3 Equalizer Filter Types

3.4.3.1 Transversal Equalizer

A training sequence used for equalization is often chosen to be a noise-like sequence, "rich" in spectral content, which is needed to estimate the channel frequency response. In the simplest sense, training might consist of sending a single narrow pulse (approximately an ideal impulse) and thereby learning the impulse response of the channel. In practice, a pseudonoise (PN) signal is preferred over a single pulse for the training sequence because the PN signal has larger average power and hence larger SNR for the same peak transmitted power. For describing the transversal filter, consider that a single pulse was transmitted over a system designated to have a raised-cosine transfer function $H_{RC}(f) = H_t(f)\, H_r(f)$. Also consider that the channel induces ISI, so that the received demodulated pulse exhibits distortion, as shown in Figure 3.25, such that the pulse sidelobes do not go through zero at sample times adjacent to the mainlobe of the pulse. The distortion can be viewed as positive or negative echoes occurring both before and after the mainlobe. To achieve the desired raised-cosine transfer function, the equalizing filter should have a frequency response $H_e(f)$, as shown in Equation (3.85), such that the actual channel response when multiplied by $H_e(f)$ yields $H_{RC}(f)$. In other words, we would like the equalizing filter to generate a set of canceling echoes.

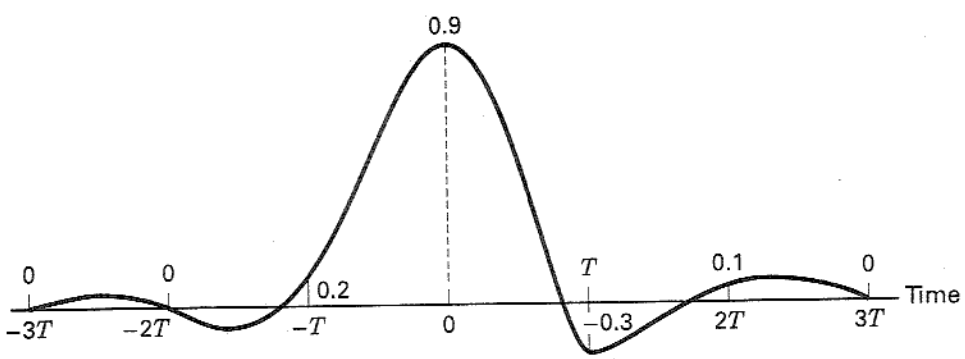

Figure 3.25 Received pulse exhibiting distortion.