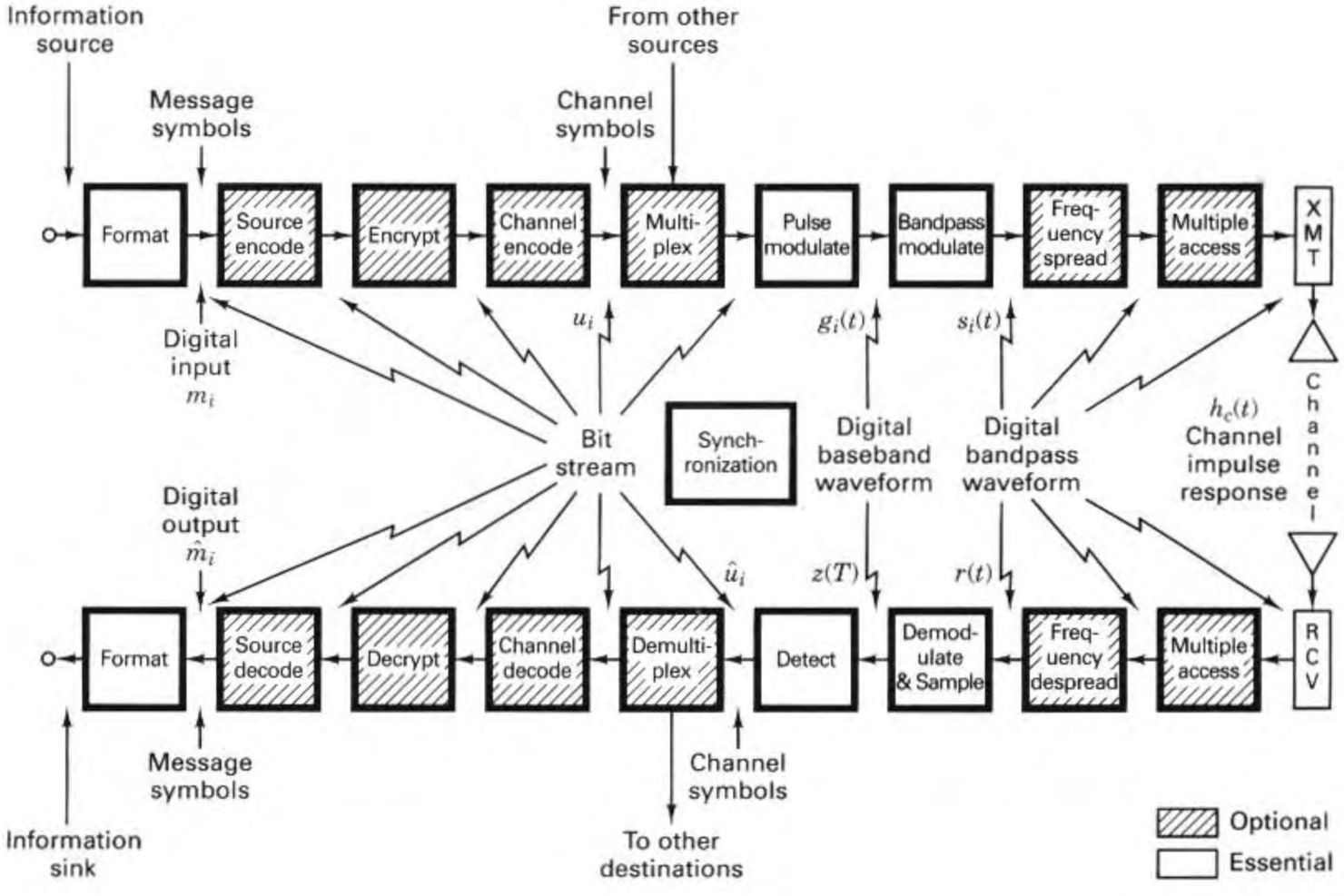
1 SIGNALS AND SPECTRA 1

CHAPTER 1

Signals and Spectra



This book presents the ideas and techniques fundamental to digital communication systems. Emphasis is placed on system design goals and on the need for trade-offs among basic system parameters such as signal-to-noise ratio (SNR), probability of error, and bandwidth expenditure. We shall deal with the transmission of information (voice, video, or data) over a path (channel) that may consist of wires, waveguides, or space.

Digital communication systems are becoming increasingly attractive because of the ever-growing demand for data communication and because digital transmission offers data processing options and flexibilities not available with analog transmission. In this book, a digital system is often treated in the context of a satellite communications link. Sometimes the treatment is in the context of a mobile radio system, in which case signal transmission typically suffers from a phenomenon called *fading*. In general, the task of characterizing and mitigating the degradation effects of a fading channel is more challenging than performing similar tasks for a nonfading channel.

The principal feature of a digital communication system (DCS) is that during a finite interval of time, it sends a waveform from a finite set of possible waveforms, in contrast to an analog communication system, which sends a waveform from an infinite variety of waveform shapes with theoretically infinite resolution. In a DCS, the objective at the receiver is *not* to reproduce a transmitted waveform with precision; instead, the objective is to determine from a noise-perturbed signal which waveform from the finite set of waveforms was sent by the transmitter. An important measure of system performance in a DCS is the probability of error (P_E).

1.1.1 Why Digital?

Why are communication systems, military and commercial alike, "going digital"? There are many reasons. The primary advantage is the ease with which digital signals, compared with analog signals, are regenerated. Figure 1.1 illustrates an ideal binary digital pulse propagating along a transmission line. The shape of the waveform is affected by two basic mechanisms: (1) as all transmission lines and circuits have some nonideal frequency transfer function, there is a distorting effect on the ideal pulse; and (2) unwanted electrical noise or other interference further distorts the pulse waveform. Both of these mechanisms cause the pulse shape to degrade as a function of line length, as shown in Figure 1.1. During the time that the transmitted pulse can still be reliably identified (before it is degraded to an ambiguous state), the pulse is amplified by a digital amplifier that recovers its original ideal shape. The pulse is thus "reborn" or regenerated. Circuits that perform this function at regular intervals along a transmission system are called *regenerative repeaters.*

Digital circuits are less subject to distortion and interference than are analog circuits. Because binary digital circuits operate in one of two states—fully on or fully off—to be meaningful, a disturbance must be large enough to change the circuit operating point from one state to the other. Such two-state operation facilitates signal regeneration and thus prevents noise and other disturbances from accumulating in transmission. Analog signals, however, are *not* two-state signals; they can take an *infinite variety* of shapes. With analog circuits, even a small disturbance can render the reproduced waveform unacceptably distorted. Once the analog signal is distorted, the distortion cannot be removed by amplification. Because accumulated noise is irrevocably bound to analog signals, they cannot be perfectly regenerated. With digital techniques, extremely low error rates producing

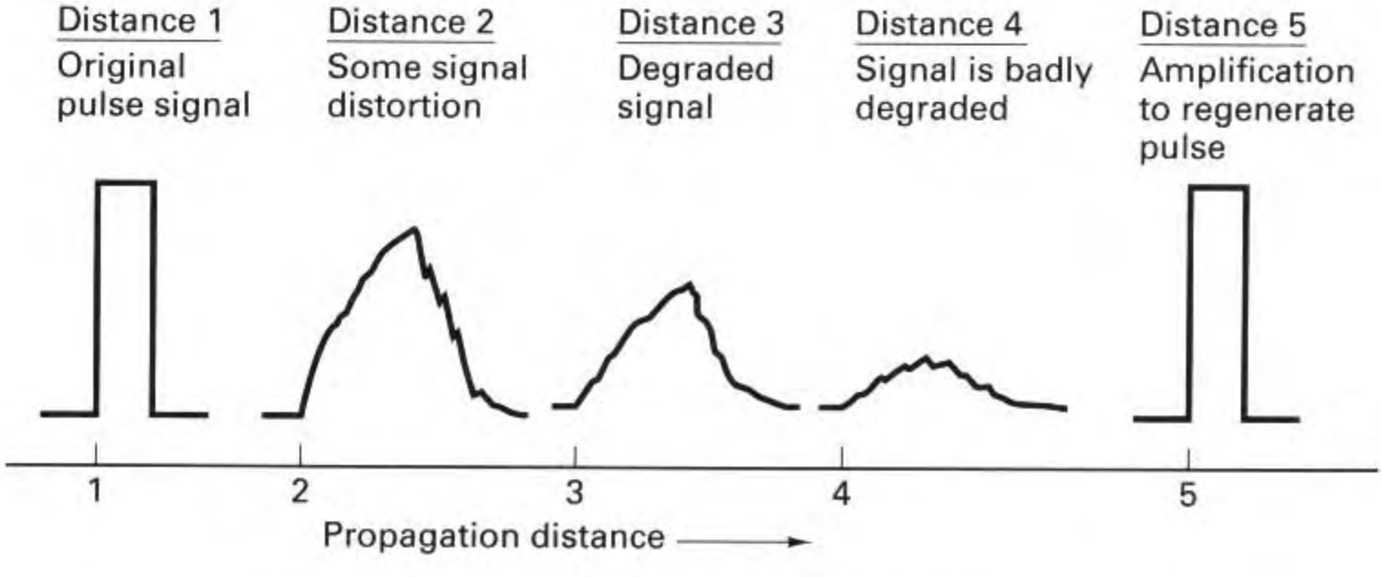


Figure 1.1 Pulse degradation and regeneration.

high signal fidelity are possible through error detection and correction but similar procedures are not available with analog.

There are other important advantages to digital communications. Digital circuits are *more reliable* and can be produced at a lower cost than analog circuits. Also, digital hardware lends itself to *more flexible* implementation than analog hardware [e.g., microprocessors, digital switching, and large-scale integrated (LSI) circuits]. The combining of digital signals using time-division multiplexing (TDM) is *simpler* than the combining of analog signals using frequency-division multiplexing (FDM). Different types of digital signals (data, telegraph, telephone, television) can be treated as identical signals in transmission and switching—*a bit is a bit.* Also, for convenient switching, digital messages can be handled in autonomous groups called *packets.* Digital techniques lend themselves naturally to signal processing functions that protect against interference and jamming, or that provide encryption and privacy. (Such techniques are discussed in Chapters 12 and 14, respectively.) Also, much data communication is from computer to computer, or from digital instruments or terminal to computer. Such digital terminations are naturally best served by digital communication links.

What are the costs associated with the beneficial attributes of digital communication systems? Digital systems tend to be very signal-processing intensive compared with analog. Also, digital systems need to allocate a significant share of their resources to the task of synchronization at various levels. (See Chapter 10.) With analog systems, on the other hand, synchronization often is accomplished more easily. One disadvantage of a digital communication system is *nongraceful degradation.* When the signal-to-noise ratio drops below a certain threshold, the quality of service can change suddenly from very good to very poor. In contrast, most analog communication systems degrade more gracefully.

1.1.2 Typical Block Diagram and Transformations

The functional block diagram shown in Figure 1.2 illustrates the signal flow and the signal-processing steps through a typical digital communication system (DCS). This figure can serve as a kind of road map, guiding the reader through the chapters of this book. The upper blocks—format, source encode, encrypt, channel encode, multiplex, pulse modulate, bandpass modulate, frequency spread, and multiple access—denote signal transformations from the source to the transmitter (XMT). The lower blocks denote signal transformations from the receiver (RCV) to the sink, essentially reversing the signal processing steps performed by the upper blocks. The *modulate* and *demodulate/detect* blocks together are called a *modem.* The term "modem" often encompasses several of the signal processing steps shown in Figure 1.2; when this is the case, the modem can be thought of as the "brains" of the system. The transmitter and receiver can be thought of as the "muscles" of the system. For wireless applications, the transmitter consists of a frequency up-conversion stage to a radio frequency (RF), a high-power amplifier, and an antenna. The receiver portion consists of an antenna and a low-noise amplifier (LNA). Frequency down-conversion is performed in the front end of the receiver and/or the demodulator.

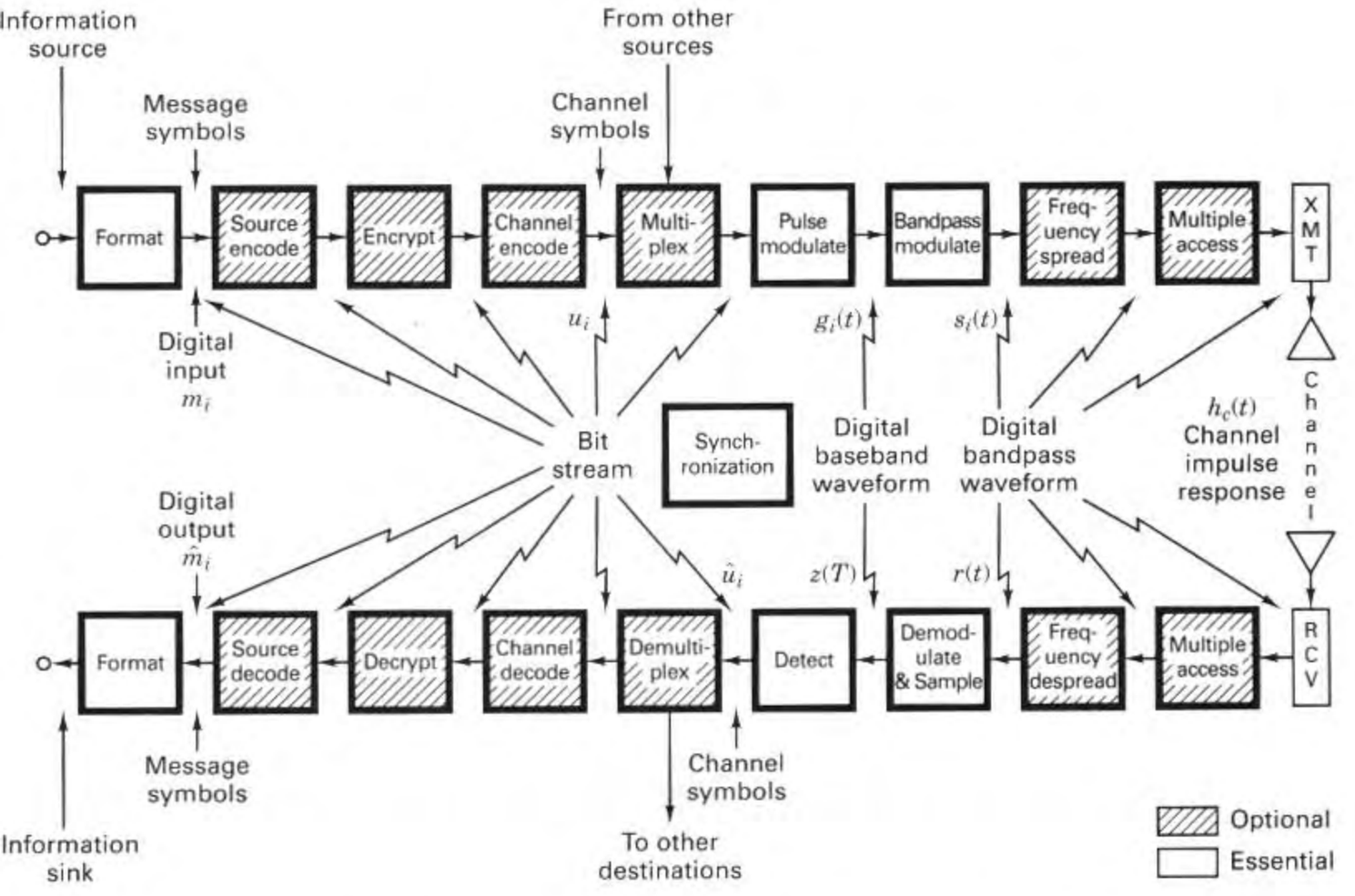


Figure 1.2 Block diagram of a typical digital communication system.

Figure 1.2 illustrates a kind of reciprocity between the blocks in the upper transmitter part of the figure and those in the lower receiver part. The signal processing steps that take place in the transmitter are, for the most part, reversed in the receiver. In Figure 1.2, the input information source is converted to binary digits (*bits*); the bits are then grouped to form *digital messages* or *message symbols.* Each such symbol (m_i, where $i = 1, \ldots, M$) can be regarded as a member of a *finite alphabet* set containing M members. Thus, for $M = 2$, the message symbol m_i is binary (meaning that it constitutes just a single bit). Even though binary symbols fall within the general definition of M-ary, nevertheless the name M-ary is usually applied to those cases where $M > 2$; hence, such symbols are each made up of a sequence of two or more bits. (Compare such a finite alphabet in a DCS with an analog system, where the message waveform is typically a member of an infinite set of possible waveforms.) For systems that use *channel coding* (error correction coding), a sequence of message symbols becomes transformed to a sequence of *channel symbols* (code symbols), where each channel symbol is denoted u_i. Because a message symbol or a channel symbol can consist of a single bit or a grouping of bits, a sequence of such symbols is also described as a *bit stream,* as shown in Figure 1.2.

Consider the key signal processing blocks shown in Figure 1.2; only formatting, modulation, demodulation/detection, and synchronization are essential for a DCS. *Formatting* transforms the source information into bits, thus assuring com-

patibility between the information and the signal processing within the DCS. From this point in the figure up to the pulse-modulation block, the information remains in the form of a *bit stream.* Modulation is the process by which message symbols or channel symbols (when channel coding is used) are converted to *waveforms* that are compatible with the requirements imposed by the transmission channel. *Pulse modulation* is an essential step because each symbol to be transmitted must first be transformed from a binary representation (voltage levels representing binary ones and zeros) to a *baseband* waveform. The term baseband refers to a signal whose spectrum extends from (or near) dc up to some finite value, usually less than a few megahertz. The pulse-modulation block usually includes filtering for minimizing the transmission bandwidth. When pulse modulation is applied to binary symbols, the resulting binary waveform is called a pulse-code-modulation (PCM) waveform. There are several types of PCM waveforms (described in Chapter 2); in telephone applications, these waveforms are often called *line codes.* When pulse modulation is applied to nonbinary symbols, the resulting waveform is called an M-ary pulse-modulation waveform. There are several types of such waveforms, and they too are described in Chapter 2, where the one called *pulse-amplitude modulation* (PAM) is emphasized. After pulse modulation, each message symbol or channel symbol takes the form of a baseband waveform $g_i(t)$, where $i = 1, \ldots, M$. In any electronic implementation, the bit stream, prior to pulse-modulation, is represented with voltage levels. One might wonder why there is a separate block for pulse modulation when in fact different voltage levels for binary ones and zeros can be viewed as impulses or as ideal rectangular pulses, each pulse occupying one bit time. There are two important differences between such voltage levels and the baseband waveforms used for modulation. First, the pulse-modulation block allows for a variety of binary and M-ary pulse-waveform types. Section 2.8.2 describes the different useful attributes of these types of waveforms. Second, the filtering within the pulse-modulation block yields pulses that occupy more than just one-bit time. Filtering yields pulses that are spread in time, thus the pulses are "smeared" into neighboring bit-times. This filtering is sometimes referred to as pulse shaping; it is used to contain the transmission bandwidth within some desired spectral region.

For an application involving RF transmission, the next important step is *bandpass modulation;* it is required whenever the transmission medium will not support the propagation of pulse-like waveforms. For such cases, the medium requires a bandpass waveform $s_i(t)$, where $i = 1, \ldots, M$. The term *bandpass* is used to indicate that the baseband waveform $g_i(t)$ is frequency translated by a carrier wave to a frequency that is much larger than the spectral content of $g_i(t)$. As $s_i(t)$ propagates over the channel, it is impacted by the channel characteristics, which can be described in terms of the channel's *impulse response* $h_c(t)$ (see Section 1.6.1). Also, at various points along the signal route, additive random noise distorts the received signal $r(t)$, so that its reception must be termed a corrupted version of the signal $s_i(t)$ that was launched at the transmitter. The received signal $r(t)$ can be expressed as

$$r(t) = s_i(t) * h_c(t) + n(t) \qquad i = 1, \ldots, M \tag{1.1}$$

where $*$ represents a convolution operation (see Appendix A), and $n(t)$ represents a noise process (see Section 1.5.5).

In the reverse direction, the receiver front end and/or the demodulator provides frequency down-conversion for each bandpass waveform $r(t)$. The demodulator restores $r(t)$ to an optimally shaped baseband pulse $z(t)$ in preparation for detection. Typically, there can be several filters associated with the receiver and demodulator—filtering to remove unwanted high frequency terms (in the frequency down-conversion of bandpass waveforms), and filtering for pulse shaping. Equalization can be described as a filtering option that is used in or after the demodulator to reverse any degrading effects on the signal that were caused by the channel. Equalization becomes essential whenever the impulse response of the channel, $h_c(t)$, is so poor that the received signal is badly distorted. An equalizer is implemented to compensate for (i.e., remove or diminish) any signal distortion caused by a nonideal $h_c(t)$. Finally, the sampling step transforms the shaped pulse $z(t)$ to a sample $z(T)$, and the detection step transforms $z(T)$ to an estimate of the channel symbol $\hat{u}_i$ or an estimate of the message symbol $\hat{m}_i$ (if there is no channel coding). Some authors use the terms "demodulation" and "detection" interchangeably. However, in this book, *demodulation* is defined as recovery of a waveform (baseband pulse), and *detection* is defined as decision-making regarding the digital meaning of that waveform.

The other signal processing steps within the modem are design options for specific system needs. *Source coding* produces analog-to-digital (A/D) conversion (for analog sources) *and* removes redundant (unneeded) information. Note that a typical DCS would either use the *source coding* option (for both digitizing and compressing the source information), or it would use the simpler *formatting* transformation (for digitizing alone). A system would not use both source coding and formatting, because the former already includes the essential step of digitizing the information. Encryption, which is used to provide communication privacy, prevents unauthorized users from understanding messages and from injecting false messages into the system. *Channel coding,* for a given data rate, can reduce the probability of error, P_E, or reduce the required signal-to-noise ratio to achieve a desired P_E at the expense of transmission bandwidth or decoder complexity. *Multiplexing* and *multiple-access procedures* combine signals that might have different characteristics or might originate from different sources, so that they can share a portion of the communications resource (e.g., spectrum, time). Frequency spreading can produce a signal that is relatively invulnerable to interference (both natural and intentional) and can be used to enhance the privacy of the communicators. It is also a valuable technique used for multiple access.

The signal processing blocks shown in Figure 1.2 represent a typical arrangement; however, these blocks are sometimes implemented in a different order. For example, multiplexing can take place prior to channel encoding, *or* prior to modulation, *or*—with a two-step modulation process (subcarrier and carrier)—it can be performed between the two modulation steps. Similarly, frequency spreading can take place at various locations along the upper portion of Figure 1.2; its precise location depends on the particular technique used. Synchronization and its key element, a clock signal, is involved in the control of all signal processing within the

DCS. For simplicity, the synchronization block in Figure 1.2 is drawn without any connecting lines, when in fact it actually plays a role in regulating the operation of almost every block shown in the figure.

Figure 1.3 shows the basic signal processing functions, which may be viewed as transformations, classified into the following nine groups:

1. Formatting and source coding
2. Baseband signaling
3. Bandpass signaling
4. Equalization
5. Channel coding
6. Multiplexing and multiple access
7. Spreading
8. Encryption
9. Synchronization

Although this organization has some inherent overlap, it provides a useful structure for the book. Beginning with Chapter 2, the nine basic transformations are considered individually. In Chapter 2, the basic formatting techniques for transforming the source information into message symbols are discussed, as well as the selection of baseband pulse waveforms and pulse filtering for making the message symbols compatible with baseband transmission. The reverse steps of demodulation, equalization, sampling, and detection are described in Chapter 3. Formatting and source coding are similar processes, in that they both involve data digitization. However, the term "source coding" has taken on the connotation of data compression in addition to digitization; it is treated later (in Chapter 13), as a special case of formatting.

In Figure 1.3, the *Baseband Signaling* block contains a list of binary choices under the heading of PCM waveforms or line codes. In this block, a nonbinary category of waveforms called M-ary pulse modulation is also listed. Another transformation in Figure 1.3, labeled *Bandpass Signaling* is partitioned into two basic blocks, coherent and noncoherent. Demodulation is typically accomplished with the aid of *reference* waveforms. When the references used are a measure of all the signal attributes (particularly phase), the process is termed *coherent;* when phase information is not used, the process is termed *noncoherent.* Both techniques are detailed in Chapter 4.

Chapter 5 is devoted to *link analysis.* Of the many specifications, analyses, and tabulations that support a developing communication system, link analysis stands out in its ability to provide overall system insight. In Chapter 5 we bring together all the link fundamentals that are essential for the analysis of most communication systems.

Channel coding deals with the techniques used to enhance digital signals so that they are less vulnerable to such channel impairments as noise, fading, and jamming. In Figure 1.3 channel coding is partitioned into two blocks, waveform coding

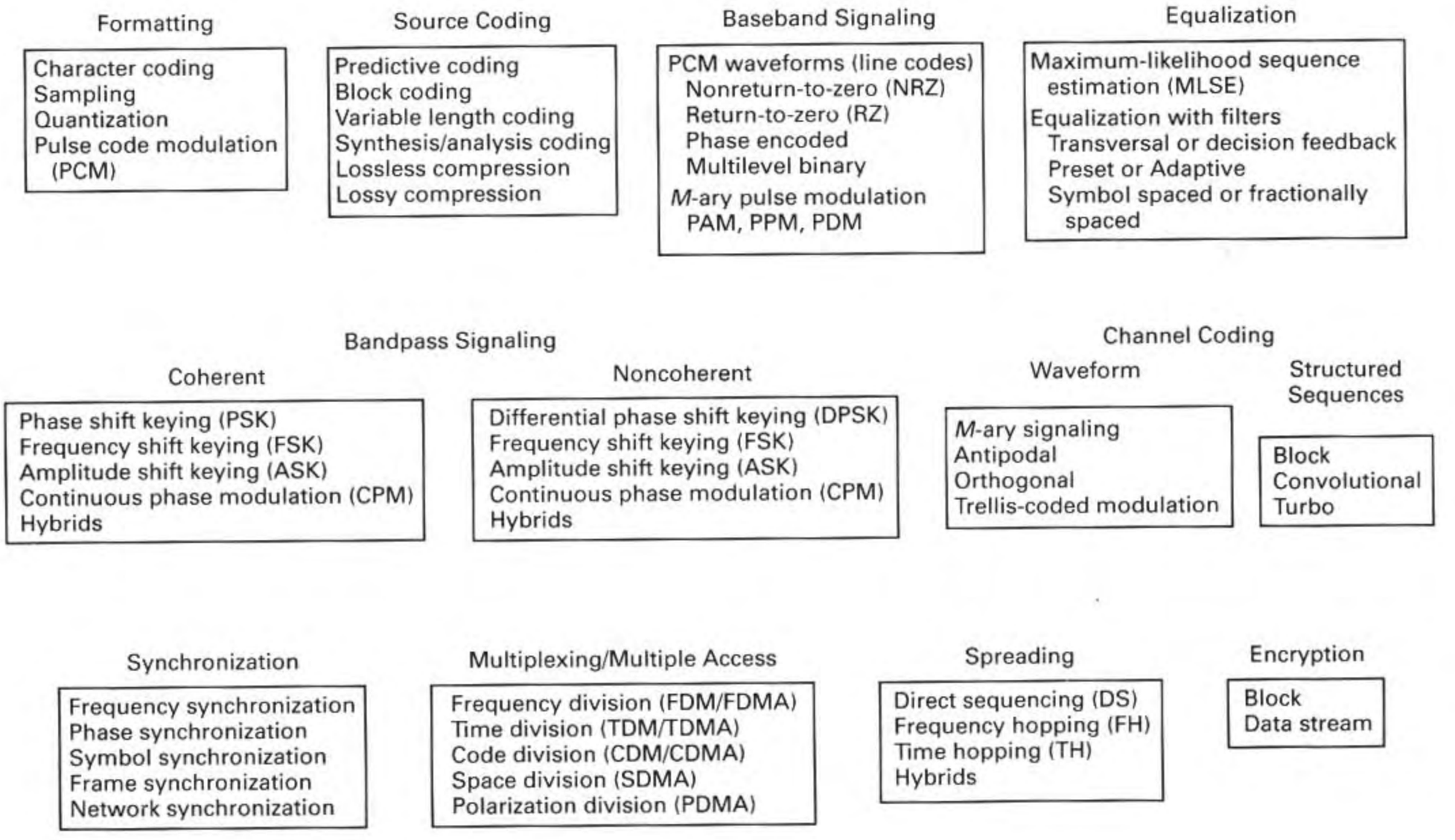


Figure 1.3 Basic digital communication transformations.

and structured sequences. *Waveform coding* involves the use of new waveforms, yielding improved detection performance over that of the original waveforms. *Structured sequences* involve the use of redundant bits to determine whether or not an error has occurred due to noise on the channel. One of these techniques, known as *automatic repeat request* (ARQ), simply recognizes the occurrence of an error and requests that the sender retransmit the message; other techniques, known as *forward error correction* (FEC), are capable of automatically correcting the errors (within specified limitations). Under the heading of structured sequences, we shall discuss three prevalent techniques—block, convolutional, and turbo coding. In Chapter 6, we primarily consider *linear block coding.* In Chapter 7 we consider *convolutional coding,* Viterbi decoding (and other decoding algorithms), and hard versus soft decoding procedures. Chapter 8 treats concatenated coding, which has led to the class of codes known as *turbo* codes, and it also examines the details of *Reed-Solomon* codes.

In Chapter 9 we summarize the design goals for a communication system and present various modulation and coding trade-offs that need to be considered in the design of a system. Theoretical limitations, such as the Nyquist criterion and the Shannon limit, are discussed. Also, *bandwidth-efficient* modulation schemes, such as trellis-coded modulation, are examined.

Chapter 10 deals with *synchronization.* In digital communications, synchronization involves the estimation of both time and frequency. The subject is divided into five subcategories as shown in Figure 1.3. Coherent systems need to synchronize their frequency reference with the carrier (and possibly subcarrier) in both frequency and phase. For noncoherent systems, phase synchronization is not needed. The fundamental time-synchronization process is symbol synchronization (or bit synchronization for binary symbols). The demodulator and detector need to know when to start and end the process of symbol detection and bit detection; a timing error will degrade detection performance. The next time-synchronization level, frame synchronization, allows the reconstruction of the message. Finally, network synchronization allows coordination with other users so resources may be used efficiently. In Chapter 10, we are concerned with the alignment of the timing of spatially separated periodic processes.

Chapter 11 deals with *multiplexing* and *multiple access.* The two terms mean very similar things. Both involve the idea of resource sharing. The main difference between the two is that multiplexing takes place locally (e.g., on a printed circuit board, within an assembly, or even within a facility), and multiple access takes place remotely (e.g., multiple users need to share the use of a satellite transponder). Multiplexing involves an algorithm that is known a priori; usually, it is hardwired into the system. Multiple access, on the other hand, is generally adaptive, and may require some overhead to enable the algorithm to operate. In Chapter 11, we discuss the classical ways of sharing a communications resource: frequency division, time division, and code division. Also, some of the multiple-access techniques that have emerged as a result of satellite communications are considered.

Chapter 12 introduces a transformation originally developed for military communications called *spreading.* The chapter deals with the spread spectrum techniques that are important for achieving interference protection and privacy.

Signals can be spread in frequency, in time, or in both frequency and time. This chapter primarily deals with frequency spreading. The chapter also illustrates how frequency-spreading techniques are used to share the bandwidth-limited resource in commercial cellular telephony.

Chapter 13 treats *source coding,* which involves the efficient description of source information. It deals with the process of compactly describing a signal to within a specified fidelity criterion. Source coding can be applied to digital or analog signals; by reducing data redundancy, source codes can reduce a system's data rate. Thus, the main advantage of source coding is to decrease the amount of required system resources (e.g., bandwidth).

Chapter 14 deals with *encryption* and *decryption,* the basic goals of which are communication privacy and authentication. Maintaining privacy means preventing unauthorized persons from extracting information (eavesdropping) from the channel. Establishing authentication means preventing unauthorized persons from injecting spurious signals (spoofing) into the channel. In this chapter we highlight the data encryption standard (DES) and the basic ideas regarding a class of encryption systems called *public key cryptosystems.* We also examine the novel scheme of Pretty Good Privacy (PGP) which is an important file-encryption method for sending data via electronic mail.

The final chapter of the book, Chapter 15, deals with fading channels. In it, we address fading that affects mobile systems such as cellular and personal communication systems (PCS). The chapter itemizes the fundamental fading manifestations, types of degradation, and methods to mitigate the degradation. Two particular mitigation techniques are examined: the Viterbi equalizer implemented in the Global System for Mobile Communication (GSM), and the Rake receiver used in CDMA systems.

1.1.3 Basic Digital Communication Nomenclature

The following are some of the basic digital signal nomenclature that frequently appears in digital communication literature:

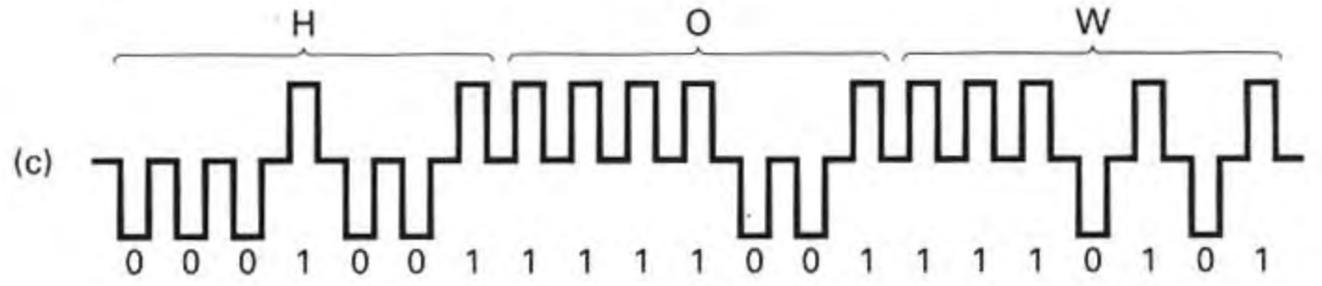
Information source. This is the device producing information to be communicated by means of the DCS. Information sources can be *analog* or *discrete.* The output of an analog source can have any value in a continuous range of amplitudes, whereas the output of a discrete information source takes its value from a finite set. Analog information sources can be transformed into digital sources through the use of *sampling* and *quantization.* Sampling and quantization techniques called formatting and source coding (see Figure 1.3) are described in Chapters 2 and 13.

Textual message. This is a sequence of characters. (See Figure 1.4a.) For digital transmission, the message will be a sequence of digits or symbols from a finite symbol set or alphabet.

Character. A character is a member of an alphabet or set of symbols. (See Figure 1.4b.) Characters may be mapped into a sequence of binary digits.

(a) HOW ARE YOU?
OK
$9, 567, 216.73

(b) A
9
&



(d) 1 Binary symbol ($k = 1, M = 2$)
10 Quaternary symbol ($k = 2, M = 4$)
011 8-ary symbol ($k = 3, M = 8$)

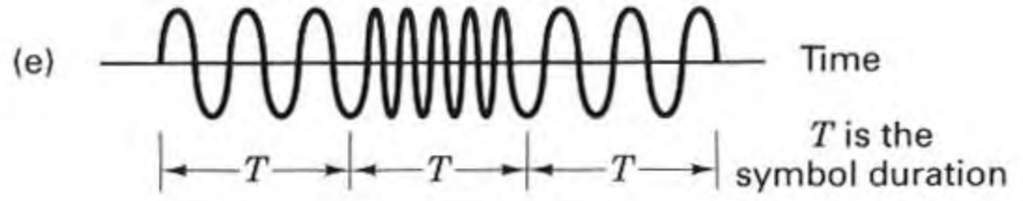


Figure 1.4 Nomenclature examples. (a) Textual messages. (b) Characters. (c) Bit stream (7-bit ASCII). (d) Symbols m_i, $i = 1, \ldots, M$, $M = 2^k$. (e) Bandpass digital waveform $s_i(t)$, $i = 1, \ldots, M$.

There are several standardized codes used for character encoding, including the American Standard Code for Information Interchange (ASCII), Extended Binary Coded Decimal Interchange Code (EBCDIC), Hollerith, Baudot, Murray, and Morse.

Binary digit (bit). This is the fundamental information unit for all digital systems. The term *bit* also is used as a unit of information content, as described in Chapter 9.

Bit stream. This is a sequence of binary digits (ones and zeros). A bit stream is often termed a *baseband* signal, which implies that its spectral content extends from (or near) dc up to some finite value, usually less than a few megahertz. In Figure 1.4c, the message, "HOW," is represented with the 7-bit ASCII character code, where the bit stream is shown by using a convenient picture of 2-level pulses. The sequence of pulses is drawn using very stylized (ideal-rectangular) shapes with spaces between successive pulses. In a real system, the pulses would never appear as they are depicted here, because such spaces would serve no useful purpose. For a given bit rate, the spaces would increase the bandwidth needed for transmission; or, for a

given bandwidth, they would increase the time delay needed to receive the message.

Symbol (digital message). A symbol is a group of k bits considered as a unit. We refer to this unit as a *message symbol* m_i ($i = 1, \ldots, M$) from a finite symbol set or alphabet. (See Figure 1.4d.) The size of the alphabet, M, is $M = 2^k$, where k is the number of bits in the symbol. For *baseband* transmission, each m_i symbol will be represented by one of a set of baseband pulse waveforms $g_1(t), g_2(t), \ldots, g_M(t)$. When transmitting a sequence of such pulses, the unit *Baud* is sometimes used to express pulse rate (symbol rate). For typical *bandpass* transmission, each $g_i(t)$ pulse will then be represented by one of a set of bandpass waveforms $s_1(t), s_2(t), \ldots, s_M(t)$. Thus, for wireless systems, the symbol m_i is sent by transmitting the digital waveform $s_i(t)$ for T seconds, the symbol-time duration. The next symbol is sent during the next time interval, T. The fact that the symbol set transmitted by the DCS is finite is a primary difference between a DCS and an analog system. The DCS receiver need only decide which of the M waveforms was transmitted; however, an analog receiver must be capable of accurately estimating a continuous range of waveforms.

Digital waveform. This is a voltage or current waveform (a pulse for baseband transmission, or a sinusoid for bandpass transmission) that represents a digital symbol. The waveform characteristics (amplitude, width, and position for pulses or amplitude, frequency, and phase for sinusoids) allow its identification as one of the symbols in the finite symbol alphabet. Figure 1.4e shows an example of a bandpass digital waveform. Even though the waveform is sinusoidal and consequently has an analog appearance, it is called a *digital waveform* because it is encoded with digital information. In the figure, during each time interval, T, a preassigned frequency indicates the value of a digit.

Data rate. This quantity in bits per second (bits/s) is given by $R = k/\mathrm{T} = (1/T) \log_2 M$ bits/s, where k bits identify a symbol from an $M = 2^k$-symbol alphabet, and T is the k-bit symbol duration.

1.1.4 Digital versus Analog Performance Criteria

A principal difference between analog and digital communication systems has to do with the way in which we evaluate their performance. Analog systems draw their waveforms from a continuum, which therefore forms an infinite set—that is, a receiver must deal with an infinite number of possible waveshapes. The figure of merit for the performance of analog communication systems is a fidelity criterion, such as signal-to-noise ratio, percent distortion, or expected mean-square error between the transmitted and received waveforms.

By contrast, a digital communication system transmits signals that represent digits. These digits form a finite set or alphabet, and the set is known a priori to the receiver. A figure of merit for digital communication systems is the probability of incorrectly detecting a digit, or the probability of error (P_E).

1.2.1 Deterministic and Random Signals

A signal can be classified as *deterministic,* meaning that there is no uncertainty with respect to its value at any time, or as *random,* meaning that there is some degree of uncertainty before the signal actually occurs. Deterministic signals or waveforms are modeled by explicit mathematical expressions, such as $x(t) = 5 \cos 10t$. For a random waveform it is *not* possible to write such an explicit expression. However, when examined over a long period, a random waveform, also referred to as a *random process,* may exhibit certain regularities that can be described in terms of probabilities and statistical averages. Such a model, in the form of a probabilistic description of the random process, is particularly useful for characterizing signals and noise in communication systems.

1.2.2 Periodic and Nonperiodic Signals

A signal $x(t)$ is called *periodic in time* if there exists a constant $T_0 > 0$ such that

$$x(t) = x(t + T_0) \qquad \text{for } -\infty < t < \infty \tag{1.2}$$

where t denotes time. The smallest value of T_0 that satisfies this condition is called the *period* of $x(t)$. The period T_0 defines the duration of one complete cycle of $x(t)$. A signal for which there is no value of T_0 that satisfies Equation (1.2) is called a *nonperiodic signal.*

1.2.3 Analog and Discrete Signals

An *analog signal* $x(t)$ is a continuous function of time; that is, $x(t)$ is uniquely defined for all t. An electrical analog signal arises when a physical waveform (e.g., speech) is converted into an electrical signal by means of a transducer. By comparison, a *discrete signal* $x(kT)$ is one that exists only at discrete times; it is characterized by a sequence of numbers defined for each time, kT, where k is an integer and T is a fixed time interval.

1.2.4 Energy and Power Signals

An electrical signal can be represented as a voltage $v(t)$ or a current $i(t)$ with instantaneous power $p(t)$ across a resistor $\mathcal{R}$ defined by

$$p(t) = \frac{v^2(t)}{\mathcal{R}} \tag{1.3a}$$

or

$$p(t) = i^2(t)\mathcal{R} \tag{1.3b}$$

In communication systems, power is often normalized by assuming $\mathcal{R}$ to be 1 Ω, although $\mathcal{R}$ may be another value in the actual circuit. If the actual value of the power is needed, it is obtained by "denormalization" of the normalized value. For the normalized case, Equations 1.3a and 1.3b have the same form. Therefore, regardless of whether the signal is a voltage or current waveform, the normalization convention allows us to express the instantaneous power as

$$p(t) = x^2(t) \tag{1.4}$$

where $x(t)$ is either a voltage or a current signal. The energy dissipated during the time interval $(-T/2, T/2)$ by a real signal with instantaneous power expressed by Equation (1.4) can then be written as

$$E_x^T = \int_{-T/2}^{T/2} x^2(t)\, dt \tag{1.5}$$

and the average power dissipated by the signal during the interval is

$$P_x^T = \frac{1}{T} E_x^T = \frac{1}{T} \int_{-T/2}^{T/2} x^2(t)\, dt \tag{1.6}$$

The performance of a communication system depends on the received signal *energy;* higher energy signals are detected more reliably (with fewer errors) than are lower energy signals—the received *energy does the work.* On the other hand, *power* is the *rate* at which energy is delivered. It is important for different reasons. The power determines the voltages that must be applied to a transmitter and the intensities of the electromagnetic fields that one must contend with in radio systems (i.e., fields in waveguides that connect the transmitter to the antenna, and fields around the radiating elements of the antenna).

In analyzing communication signals, it is often desirable to deal with the *waveform energy.* We classify $x(t)$ as an *energy signal* if, and only if, it has nonzero but finite energy $(0 < E_x < \infty)$ for all time, where

$$\begin{aligned} E_x &= \lim_{T \to \infty} \int_{-T/2}^{T/2} x^2(t)\, dt \\ &= \int_{-\infty}^{\infty} x^2(t)\, dt \end{aligned} \tag{1.7}$$

In the real world, we always transmit signals having finite energy $(0 < E_x < \infty)$. However, in order to describe *periodic signals,* which by definition [Equation (1.2)] exist for all time and thus have infinite energy, and in order to deal with random signals that have infinite energy, it is convenient to define a class of signals called *power signals.* A signal is defined as a power signal if, and only if, it has finite but nonzero power $(0 < P_x < \infty)$ for all time, where

$$P_x = \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} x^2(t)\, dt \tag{1.8}$$

The energy and power classifications are mutually exclusive. An energy signal has finite energy but *zero average power,* whereas a power signal has finite average power but *infinite energy.* A waveform in a system may be constrained in either its power or energy values. As a general rule, periodic signals and random signals are classified as power signals, while signals that are both deterministic and nonperiodic are classified as energy signals [1, 2].

Signal energy and power are both important parameters in specifying a communication system. The classification of a signal as either an energy signal or a power signal is a convenient model to facilitate the mathematical treatment of various signals and noise. In Section 3.1.5, these ideas are developed further, in the context of a digital communication system.

1.2.5 The Unit Impulse Function

A useful function in communication theory is the unit impulse or *Dirac delta function* $\delta(t)$. The impulse function is an abstraction—an infinitely large amplitude pulse, with zero pulse width, and unity weight (area under the pulse), concentrated at the point where its argument is zero. The unit impulse is characterized by the following relationships:

$$\int_{-\infty}^{\infty} \delta(t)\, dt = 1 \tag{1.9}$$

$$\delta(t) = 0 \qquad \text{for } t \neq 0 \tag{1.10}$$

$$\delta(t) \text{ is unbounded at } t = 0 \tag{1.11}$$

$$\int_{-\infty}^{\infty} x(t)\delta(t - t_0)\, dt = x(t_0) \tag{1.12}$$

The unit impulse function $\delta(t)$ is not a function in the usual sense. When operations involve $\delta(t)$, the convention is to interpret $\delta(t)$ as a unit-area pulse of finite amplitude and nonzero duration, after which the limit is considered as the pulse duration approaches zero. $\delta(t - t_0)$ can be depicted graphically as a spike located at $t = t_0$ with height equal to its integral or area. Thus $A\delta(t - t_0)$ with A constant represents an impulse function whose area or weight is equal to A, that is zero everywhere except at $t = t_0$.

Equation (1.12) is known as the *sifting* or *sampling property* of the unit impulse function; the unit impulse multiplier selects a sample of the function $x(t)$ evaluated at $t = t_0$.

1.3 SPECTRAL DENSITY

The *spectral density* of a signal characterizes the distribution of the signal's energy or power in the frequency domain. This concept is particularly important when considering filtering in communication systems. We need to be able to evaluate the

signal and noise at the filter output. The energy spectral density (ESD) or the power spectral density (PSD) is used in the evaluation.

1.3.1 Energy Spectral Density

The total energy of a real-valued energy signal $x(t)$, defined over the interval, $(-\infty, \infty)$, is described by Equation (1.7). Using Parseval's theorem [1], we can relate the energy of such a signal expressed in the time domain to the energy expressed in the frequency domain, as

$$E_x = \int_{-\infty}^{\infty} x^2(t)\, dt = \int_{-\infty}^{\infty} |X(f)|^2\, df \tag{1.13}$$

where $X(f)$ is the Fourier transform of the nonperiodic signal $x(t)$. (For a review of Fourier techniques, see Appendix A.) Let $\psi_x(f)$ denote the squared magnitude spectrum, defined as

$$\psi_x(f) = |X(f)|^2 \tag{1.14}$$

The quantify $\psi_x(f)$ is the waveform *energy spectral density* (ESD) of the signal $x(t)$. Therefore, from Equation (1.13), we can express the total energy of $x(t)$ by integrating the spectral density with respect to frequency:

$$E_x = \int_{-\infty}^{\infty} \psi_x(f)\, df \tag{1.15}$$

This equation states that the energy of a signal is equal to the area under the $\psi_x(f)$ versus frequency curve. Energy spectral density describes the signal energy per unit bandwidth measured in joules/hertz. There are equal energy contributions from both positive and negative frequency components, since for a real signal, $x(t)$, $|X(f)|$ is an even function of frequency. Therefore, the energy spectral density is symmetrical in frequency about the origin, and thus the total energy of the signal $x(t)$ can be expressed as

$$E_x = 2\int_{0}^{\infty} \psi_x(f)\, df \tag{1.16}$$

1.3.2 Power Spectral Density

The average power P_x of a real-valued power signal $x(t)$ is defined in Equation (1.8). If $x(t)$ is a *periodic signal* with period T_0, it is classified as a power signal. The expression for the average power of a periodic signal takes the form of Equation (1.6), where the time average is taken over the signal period T_0, as follows:

$$P_x = \frac{1}{T_0}\int_{-T_0/2}^{T_0/2} x^2(t)\, dt \tag{1.17a}$$

Parseval's theorem for a real-valued periodic signal [1] takes the form

$$P_x = \frac{1}{T_0}\int_{-T_0/2}^{T_0/2} x^2(t)\,dt = \sum_{n=-\infty}^{\infty} |c_n|^2 \tag{1.17b}$$

where the $|c_n|$ terms are the complex Fourier series coefficients of the periodic signal. (See Appendix A.)

To apply Equation (1.17b), we need only know the magnitude of the coefficients, $|c_n|$. The *power spectral density* (PSD) function $G_x(f)$ of the periodic signal $x(t)$ is a real, even, and nonnegative function of frequency that gives the distribution of the power of $x(t)$ in the frequency domain, defined as

$$G_x(f) = \sum_{n=-\infty}^{\infty} |c_n|^2\,\delta(f - nf_0) \tag{1.18}$$

Equation (1.18) defines the power spectral density of a periodic signal $x(t)$ as a succession of the weighted delta functions. Therefore, the PSD of a periodic signal is a discrete function of frequency. Using the PSD defined in Equation (1.18), we can now write the average normalized power of a real-valued signal as

$$P_x = \int_{-\infty}^{\infty} G_x(f)\,df = 2\int_{0}^{\infty} G_x(f)\,df \tag{1.19}$$

Equation (1.18) describes the PSD of periodic (power) signals only. If $x(t)$ is a nonperiodic signal it *cannot* be expressed by a Fourier series, and if it is a nonperiodic power signal (having infinite energy) it *may not* have a Fourier transform. However, we may still express the power spectral density of such signals in the *limiting sense.* If we form a *truncated version* $x_T(t)$ of the nonperiodic power signal $x(t)$ by observing it only in the interval $(-T/2, T/2)$, then $x_T(t)$ has finite energy and has a proper Fourier transform $X_T(f)$. It can be shown [2] that the power spectral density of the nonperiodic $x(t)$ can then be defined in the limit as

$$G_x(f) = \lim_{T\to\infty} \frac{1}{T}\,|X_T(f)|^2 \tag{1.20}$$

Example 1.1 Average Normalized Power

(a) Find the average normalized power in the waveform, $x(t) = A\cos 2\pi f_0 t$, using time averaging.

(b) Repeat part (a) using the summation of spectral coefficients.

Solution

(a) Using Equation (1.17a), we have

$$P_x = \frac{1}{T_0}\int_{-T_0/2}^{T_0/2} A^2\cos^2 2\pi f_0 t\;dt$$

$$= \frac{A^2}{2T_0}\int_{-T_0/2}^{T_0/2} (1 + \cos 4\pi f_0 t)\,dt$$

$$= \frac{A^2}{2T_0}(T_0) = \frac{A^2}{2}$$

(b) Using Equations (1.18) and (1.19) gives us

$$G_x(f) = \sum_{n=-\infty}^{\infty} |c_n|^2 \, \delta(f - nf_0)$$

$$\left.\begin{aligned} c_1 = c_{-1} &= \frac{A}{2} \\ c_n &= 0 \qquad \text{for } n = 0, \pm 2, \pm 3, \ldots \end{aligned}\right\} \quad \text{(see Appendix A)}$$

$$G_x(f) = \left(\frac{A}{2}\right)^2 \delta(f - f_0) + \left(\frac{A}{2}\right)^2 \delta(f + f_0)$$

$$P_x = \int_{-\infty}^{\infty} G_x(f)\, df = \frac{A^2}{2}$$

1.4 AUTOCORRELATION

1.4.1 Autocorrelation of an Energy Signal

Correlation is a matching process; *autocorrelation* refers to the matching of a signal with a delayed version of itself. The autocorrelation function of a real-valued energy signal $x(t)$ is defined as

$$R_x(\tau) = \int_{-\infty}^{\infty} x(t)x(t + \tau)\, dt \qquad \text{for } -\infty < \tau < \infty \tag{1.21}$$

The autocorrelation function $R_x(\tau)$ provides a measure of how closely the signal matches a copy of itself as the copy is shifted τ units in time. The variable τ plays the role of a scanning or searching parameter. $R_x(\tau)$ is not a function of time; it is only a function of the time difference τ between the waveform and its shifted copy.

The autocorrelative function of a real-valued *energy* signal has the following properties:

1. $R_x(\tau) = R_x(-\tau)$	symmetrical in τ about zero
2. $R_x(\tau) \le R_x(0)$ for all τ	maximum value occurs at the origin
3. $R_x(\tau) \leftrightarrow \psi_x(f)$	autocorrelation and ESD form a Fourier transform pair, as designated by the double-headed arrows
4. $R_x(0) = \int_{-\infty}^{\infty} x^2(t)\, dt$	value at the origin is equal to the energy of the signal

If items 1 through 3 are satisfied, $R_x(\tau)$ satisfies the properties of an autocorrelation function. Property 4 can be derived from property 3 and thus need not be included as a basic test.

1.4.2 Autocorrelation of a Periodic (Power) Signal

The autocorrelation function of a real-valued power signal $x(t)$ is defined as

$$R_x(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t)x(t+\tau)\,dt \qquad \text{for } -\infty < \tau < \infty \tag{1.22}$$

When the power signal $x(t)$ is periodic with period T_0, the time average in Equation (1.22) may be taken over a *single period* T_0, and the autocorrelation function can be expressed as

$$R_x(\tau) = \frac{1}{T_0} \int_{-T_0/2}^{T_0/2} x(t)x(t+\tau)\,dt \qquad \text{for } -\infty < \tau < \infty \tag{1.23}$$

The autocorrelation function of a real-valued *periodic* signal has properties similar to those of an energy signal:

1. $R_x(\tau) = R_x(-\tau)$ — symmetrical in τ about zero
2. $R_x(\tau) \le R_x(0)$ for all τ — maximum value occurs at the origin
3. $R_x(\tau) \leftrightarrow G_x(f)$ — autocorrelation and PSD form a Fourier transform pair
4. $R_x(0) = \frac{1}{T_0} \int_{-T_0/2}^{T_0/2} x^2(t)\,dt$ — value at the origin is equal to the average power of the signal

1.5 RANDOM SIGNALS

The main objective of a communication system is the transfer of information over a channel. All useful message signals appear random; that is, the receiver does not know, a priori, which of the possible message waveforms will be transmitted. Also, the noise that accompanies the message signals is due to random electrical signals. Therefore, we need to be able to form efficient descriptions of random signals.

1.5.1 Random Variables

Let a *random variable* $X(A)$ represent the functional relationship between a random event A and a real number. For notational convenience, we shall designate the random variable by X, and let the functional dependence upon A be implicit. The random variable may be discrete or continuous. The *distribution function* $F_X(x)$ of the random variable X is given by

$$F_X(x) = P(X \le x) \tag{1.24}$$

where $P(X \le x)$ is the probability that the value taken by the random variable X is less than or equal to a real number x. The distribution function $F_X(x)$ has the following properties:

1. $0 \le F_X(x) \le 1$
2. $F_X(x_1) \le F_X(x_2)$ if $x_1 \le x_2$
3. $F_X(-\infty) = 0$
4. $F_X(+\infty) = 1$

Another useful function relating to the random variable X is the *probability density function* (pdf), denoted

$$p_X(x) = \frac{dF_X(x)}{dx} \tag{1.25a}$$

As in the case of the distribution function, the pdf is a function of a real number x. The name "density function" arises from the fact that the probability of the event $x_1 \le X \le x_2$ equals

$$\begin{aligned} P(x_1 \le X \le x_2) &= P(X \le x_2) - P(X \le x_1) \\ &= F_X(x_2) - F_X(x_1) \\ &= \int_{x_1}^{x_2} p_X(x)\, dx \end{aligned} \tag{1.25b}$$

From Equation (1.25b), the probability that a random variable X has a value in some very narrow range between x and $x + \Delta x$ can be approximated as

$$P(x \le X \le x + \Delta x) \approx p_X(x)\Delta x \tag{1.25c}$$

Thus, in the limit as Δx approaches zero, we can write

$$P(X = x) = p_X(x)dx \tag{1.25d}$$

The probability density function has the following properties:

1. $p_X(x) \ge 0$.
2. $\displaystyle\int_{-\infty}^{\infty} p_X(x)\, dx = F_X(+\infty) - F_X(-\infty) = 1.$

Thus, a probability density function is always a nonnegative function with a total area of one. Throughout the book we use the designation $p_X(x)$ for the probability density function of a *continuous* random variable. For ease of notation, we will often omit the subscript X and write simply $p(x)$. We will use the designation $p(X = x_i)$ for the probability of a random variable X, where X can take on *discrete* values only.

1.5.1.1 Ensemble Averages

The *mean value* m_X, or *expected value* of a random variable X, is defined by

$$m_X = \mathbf{E}\{X\} = \int_{-\infty}^{\infty} x p_X(x)\, dx \tag{1.26}$$

where $\mathbf{E}\{\cdot\}$ is called the *expected value operator.* The *nth moment* of a probability distribution of a random variable X is defined by

$$\mathbf{E}\{X^n\} = \int_{-\infty}^{\infty} x^n p_X(x)\, dx \tag{1.27}$$

For the purposes of communication system analysis, the most important moments of X are the first two moments. Thus, $n = 1$ in Equation (1.27) gives m_X as discussed above, whereas $n = 2$ gives the mean-square value of X, as follows:

$$\mathbf{E}\{X^2\} = \int_{-\infty}^{\infty} x^2 p_X(x)\, dx \tag{1.28}$$

We can also define *central moments,* which are the moments of the difference between X and m_X. The second central moment, called the *variance* of X, is defined as

$$\text{var}\,(X) = \mathbf{E}\{X - m_X)^2\} = \int_{-\infty}^{\infty} (x - m_X)^2 p_X(x)\, dx \tag{1.29}$$

The variance of X is also denoted as σ_X^2, and its square root, σ_X, is called the *standard deviation* of X. Variance is a measure of the "randomness" of the random variable X. By specifying the variance of a random variable, we are constraining the width of its probability density function. The variance and the mean-square value are related by

$$\begin{aligned}\sigma_X^2 &= \mathbf{E}\{X^2 - 2m_X X + m_X^2\} \\ &= \mathbf{E}\{X^2\} - 2m_X\mathbf{E}\{X\} + m_X^2 \\ &= \mathbf{E}\{X^2\} - m_X^2\end{aligned}$$

Thus, the variance is equal to the difference between the mean-square value and the square of the mean.

1.5.2 Random Processes

A random process $X(A, t)$ can be viewed as a function of two variables: *an event A* and *time.* Figure 1.5 illustrates a random process. In the figure there are N *sample functions* of time, $\{X_j(t)\}$. Each of the sample functions can be regarded as the output of a different noise generator. For a specific event A_j, we have a single time function $X(A_j, t) = X_j(t)$ (i.e., a sample function). The totality of all sample functions is called an *ensemble.* For a specific time t_k, $X(A, t_k)$ is a *random variable* $X(t_k)$ whose value depends on the event. Finally, for a specific event, $A = A_j$ and a specific time $t = t_k$, $X(A_j, t_k)$ is simply a *number.* For notational convenience we shall designate the random process by $X(t)$, and let the functional dependence upon A be implicit.

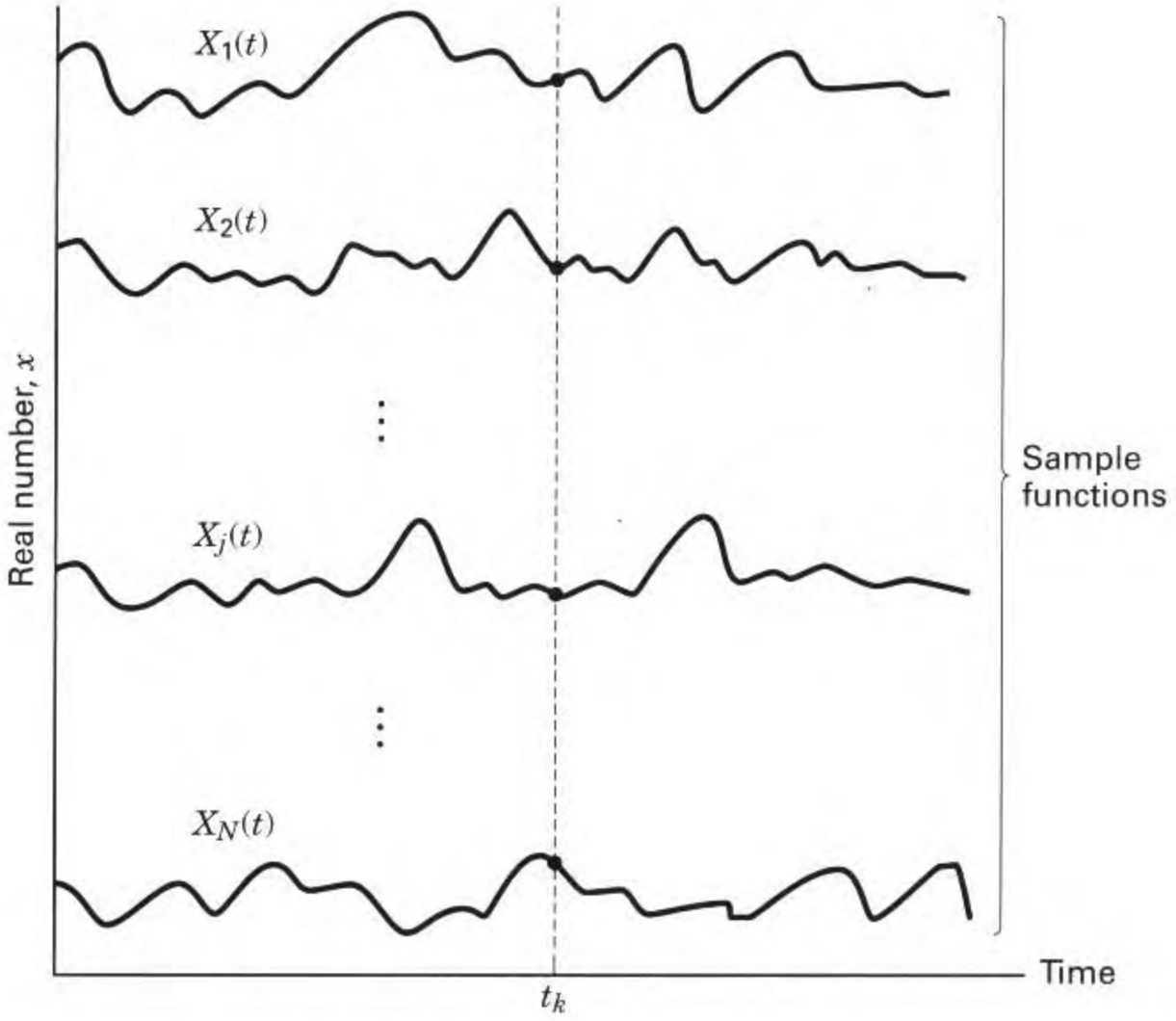


Figure 1.5 Random noise process.

1.5.2.1 Statistical Averages of a Random Process

Because the value of a random process at any future time is unknown (since the identity of the event A is unknown), a random process whose distribution functions are continuous can be described statistically with a probability density function (pdf). In general, the form of the pdf of a random process will be different for different times. In most situations it is not practical to determine empirically the probability distribution of a random process. However, a partial description consisting of the mean and autocorrelation function are often adequate for the needs of communication systems. We define the mean of the random process $X(t)$ as

$$\mathbf{E}\{X(t_k)\} = \int_{-\infty}^{\infty} x p_{X_k}(x)\, dx = m_X(t_k) \tag{1.30}$$

where $X(t_k)$ is the random variable obtained by observing the random process at time t_k and the pdf of $X(t_k)$, the density over the ensemble of events at time t_k, is designated $p_{X_k}(x)$.

We define the autocorrelation function of the random process $X(t)$ to be a function of two variables, t_1 and t_2, given by

$$R_X(t_1, t_2) = \mathbf{E}\{X(t_1)X(t_2)\} \tag{1.31}$$

where $X(t_1)$ and $X(t_2)$ are random variables obtained by observing $X(t)$ at times t_1 and t_2, respectively. The autocorrelation function is a measure of the degree to which two time samples of the same random process are related.

1.5.2.2 Stationarity

A random process $X(t)$ is said to be *stationary* in the *strict sense* if none of its statistics are affected by a shift in the time origin. A random process is said to be *wide-sense stationary* (WSS) if two of its statistics, its mean and autocorrelation function, do not vary with a shift in the time origin. Thus, a process is WSS if

$$\mathbf{E}\{X(t)\} = m_X = \text{a constant} \tag{1.32}$$

and

$$R_X(t_1, t_2) = R_X(t_1 - t_2) \tag{1.33}$$

Strict-sense stationary implies wide-sense stationary, but not vice versa. Most of the useful results in communication theory are predicated on random information signals and noise being wide-sense stationary. From a practical point of view, it is not necessary for a random process to be stationary for all time but only for some observation interval of interest.

For stationary processes, the autocorrelation function in Equation (1.33) does not depend on time but only on the difference between t_1 and t_2. That is, all pairs of values of $X(t)$ at points in time separated by $\tau = t_1 - t_2$ have the same correlation value. Thus, for stationary systems, we can denote $R_X(t_1, t_2)$ simply as $R_X(\tau)$.

1.5.2.3 Autocorrelation of a Wide-Sense Stationary Random Process

Just as the variance provides a measure of randomness for random variables, the autocorrelation function provides a similar measure for random processes. For a wide-sense stationary process, the autocorrelation function is only a function of the *time difference* $\tau = t_1 - t_2$; that is,

$$R_X(\tau) = \mathbf{E}\{X(t)X(t + \tau)\} \qquad \text{for } -\infty < \tau < \infty \tag{1.34}$$

For a zero mean WSS process, $R_X(\tau)$ indicates the extent to which the random values of the process separated by τ seconds in time are statistically correlated. In other words, $R_X(\tau)$ gives us an idea of the frequency response that is associated with a random process. If $R_X(\tau)$ changes slowly as τ increases from zero to some value, it indicates that, on average, sample values of $X(t)$ taken at $t = t_1$ and $t = t_1 + \tau$ are nearly the same. Thus, we would expect a frequency domain representation of $X(t)$ to contain a preponderance of low frequencies. On the other hand, if $R_X(\tau)$ decreases rapidly as τ is increased, we would expect $X(t)$ to change rapidly with time and thereby contain mostly high frequencies.

Properties of the autocorrelation function of a real-valued wide-sense stationary process are as follows:

1. $R_X(\tau) = R_X(-\tau)$	symmetrical in τ about zero
2. $R_X(\tau) \leq R_X(0)$ for all τ	maximum value occurs at the origin
3. $R_X(\tau) \leftrightarrow G_X(f)$	autocorrelation and power spectral density form a Fourier transform pair
4. $R_X(0) = \mathbf{E}\{X^2(t)\}$	value at the origin is equal to the average power of the signal

1.5.3 Time Averaging and Ergodicity

To compute m_X and $R_X(\tau)$ by ensemble averaging, we would have to average across all the sample functions of the process and would need to have complete knowledge of the first- and second-order joint probability density functions. Such knowledge is generally not available.

When a random process belongs to a special class, known as an *ergodic process,* its time averages equal its ensemble averages, and the statistical properties of the process can be determined by *time averaging over a single sample function* of the process. For a random process to be ergodic, it must be stationary in the strict sense. (The converse is not necessary.) However, for communication systems, where we are satisfied to meet the conditions of wide-sense stationarity, we are interested only in the mean and autocorrelation functions.

We can say that a random process is *ergodic in the mean* if

$$m_X = \lim_{T \to \infty} 1/T \int_{-T/2}^{T/2} X(t)\, dt \tag{1.35}$$

and it is *ergodic in the autocorrelation function* if

$$R_X(\tau) = \lim_{T \to \infty} 1/T \int_{-T/2}^{T/2} X(t)X(t + \tau)\, dt \tag{1.36}$$

Testing for the ergodicity of a random process is usually very difficult. In practice one makes an intuitive judgment as to whether it is reasonable to interchange the time and ensemble averages. A reasonable assumption in the analysis of most communication signals (in the absence of transient effects) is that the random waveforms are ergodic in the mean and the autocorrelation function. Since time averages equal ensemble averages for ergodic processes, fundamental electrical engineering parameters, such as dc value, rms value, and average power can be related to the moments of an ergodic random process. Following is a summary of these relationships:

1. The quantity $m_X = \mathbf{E}\{X(t)\}$ is equal to the dc level of the signal.
2. The quantity m_X^2 is equal to the normalized power in the dc component.
3. The second moment of $X(t)$, $\mathbf{E}\{X^2(t)\}$, is equal to the total average normalized power.

4. The quantity $\sqrt{\mathbf{E}\{X^2(t)\}}$ is equal to the root-mean-square (rms) value of the voltage or current signal.
5. The variance σ_X^2 is equal to the average normalized power in the time-varying or ac component of the signal.
6. If the process has zero mean (i.e., $m_X = m_X^2 = 0$), then $\sigma_X^2 = \mathbf{E}\{X^2\}$ and the variance is the same as the mean-square value, or the variance represents the total power in the normalized load.
7. The standard deviation σ_X is the rms value of the ac component of the signal.
8. If $m_X = 0$, then σ_X is the rms value of the signal.

1.5.4 Power Spectral Density and Autocorrelation of a Random Process

A random process $X(t)$ can generally be classified as a power signal having a power spectral density (PSD) $G_X(f)$ of the form shown in Equation (1.20). $G_X(f)$ is particularly useful in communication systems, because it describes the distribution of a signal's power in the frequency domain. The PSD enables us to evaluate the signal power that will pass through a network having known frequency characteristics. We summarize the principal features of PSD functions as follows:

1. $G_X(f) \geq 0$ and is always real valued
2. $G_X(f) = G_X(-f)$ for $X(t)$ real-valued
3. $G_X(f) \leftrightarrow R_X(\tau)$ PSD and autocorrelation form a Fourier transform pair
4. $P_X = \int_{-\infty}^{\infty} G_X(f)\, df$ relationship between average normalized power and PSD

In Figure 1.6, we present a visualization of autocorrelation and power spectral density functions. What does the term *correlation* mean? When we inquire about the correlation between two phenomena, we are asking how closely do they correspond in behavior or appearance, how well do they match one another. In mathematics, an autocorrelation function of a signal (in the time domain) describes the correspondence of the signal to itself in the following way. An exact copy of the signal is made and located in time at minus infinity. Then we move the copy an increment in the direction of positive time and ask the question, "How well do these two (the original versus the copy) match"? We move the copy another step in the positive direction and ask, "How well do they match now?" And so forth. The correlation between the two is plotted as a function of time, denoted τ, which can be thought of as a scanning parameter.

Figure 1.6a–d highlights some of these steps. Figure 1.6a illustrates a single sample waveform from a WSS random process, $X(t)$. The waveform is a binary random sequence with unit-amplitude positive and negative (bipolar) pulses. The positive and negative pulses occur with equal probability. The duration of each binary digit is T seconds, and the average or dc value of the random sequence is zero. Figure 1.6b shows the same sequence displaced τ_1 seconds in time; this sequence is therefore denoted $X(t - \tau_1)$. Let us assume that $X(t)$ is ergodic in the autocorrela-

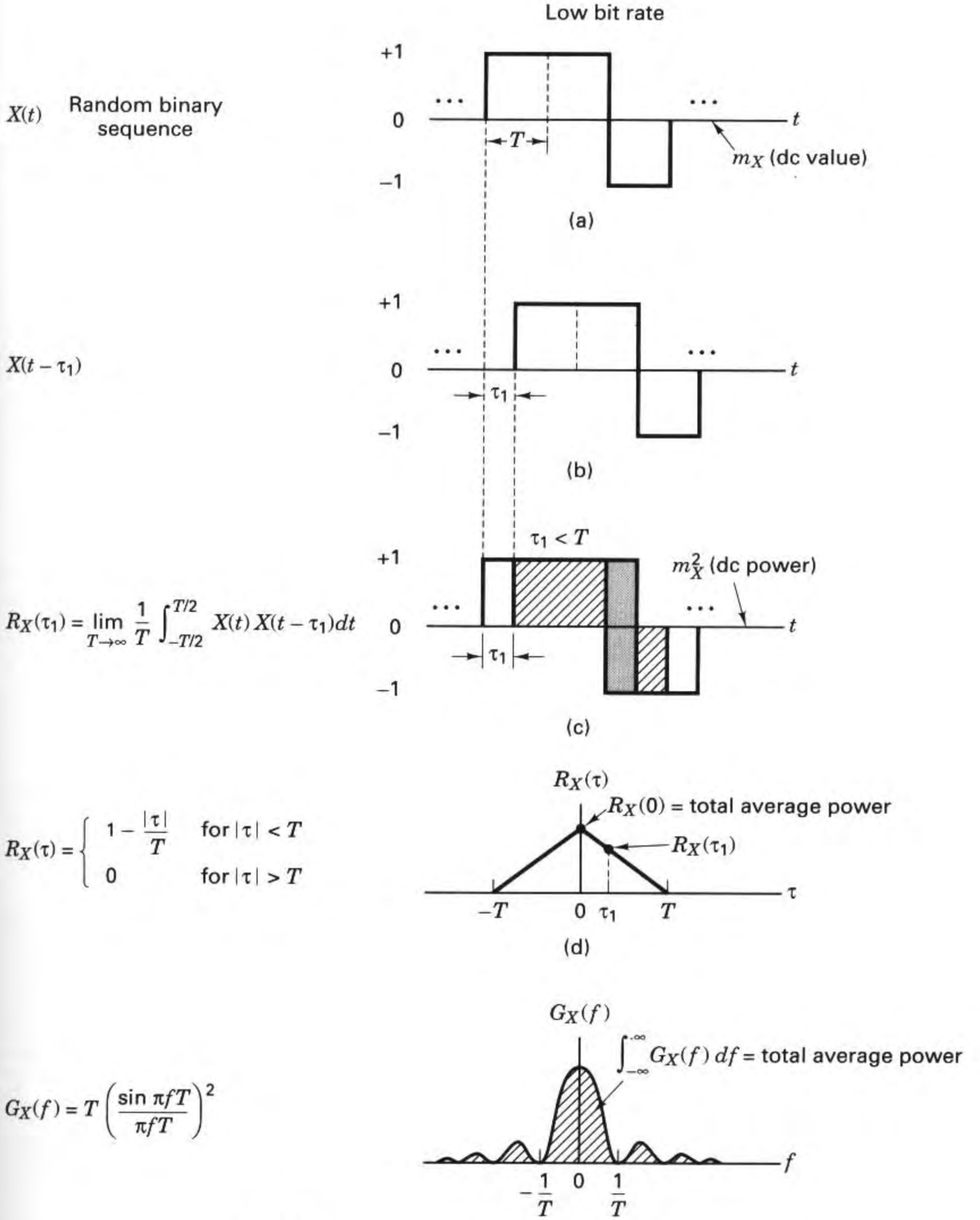
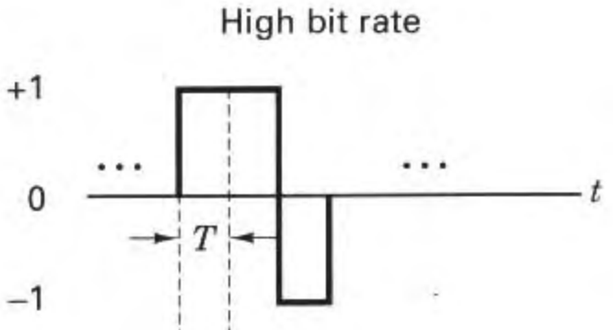


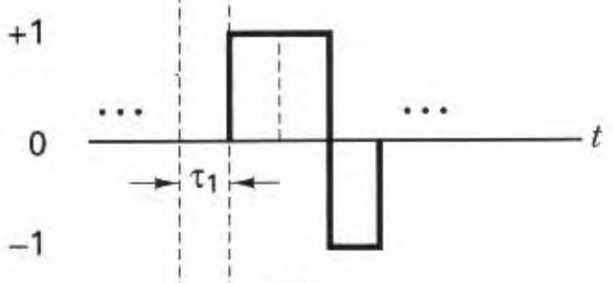
Figure 1.6 Autocorrelation and power spectral density.

$X(t)$ Random binary sequence



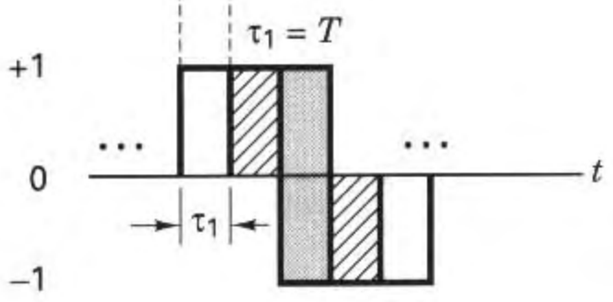
(f)

$X(t - \tau_1)$



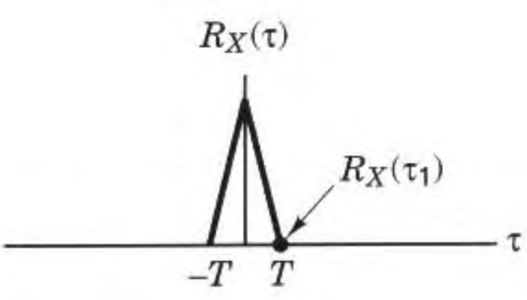
(g)

$$R_X(\tau_1) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} X(t)\, X(t - \tau_1) dt$$



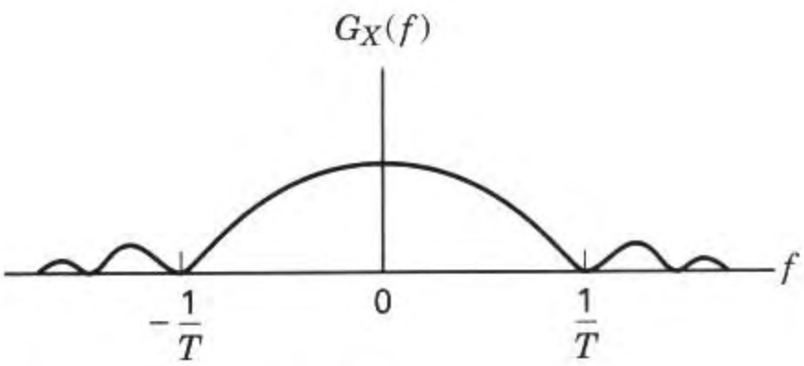
(h)

$$R_X(\tau) = \begin{cases} 1 - \dfrac{|\tau|}{T} & \text{for } |\tau| < T \\ 0 & \text{for } |\tau| > T \end{cases}$$



(i)

$$G_X(f) = T\left(\frac{\sin \pi f T}{\pi f T}\right)^2$$



(j)

Figure 1.6 continued

tion function so that we can use time averaging instead of ensemble averaging to find $R_X(\tau)$. The value of $R_X(\tau_1)$ is obtained by taking the product of the two sequences $X(t)$ and $X(t - \tau_1)$ and finding the average value using Equation (1.36). Equation (1.36) is accurate for ergodic processes *only in the limit.* However, integration over an integer number of periods can provide us with an estimate of $R_X(\tau)$. Notice that $R_X(\tau_1)$ can be obtained by a positive or negative shift of $X(t)$. Figure 1.6c illustrates such a calculation, using the single sample sequence (Figure 1.6a) and its shifted replica (Figure 1.6b). The cross-hatched areas under the product curve $X(t)X(t - \tau_1)$ contribute to positive values of the product, and the grey areas contribute to negative values. The integration of $X(t)\ X(t - \tau_1)$ over several pulse times yields a net value of area which is one point, the $R_X(\tau_1)$ point of the $R_X(\tau)$ curve. The sequences can be further shifted by $\tau_2, \tau_3, \ldots$, each shift yielding a point on the overall autocorrelation function $R_X(\tau)$ shown in Figure 1.6d. Every random sequence of bipolar pulses has an autocorrelation plot of the general shape shown in Figure 1.6d. The plot peaks at $R_X(0)$ [the best match occurs when τ equals zero, since $R(\tau) \leq R(0)$ for all τ], and it declines as τ increases. Figure 1.6d shows points corresponding to $R_X(0)$ and $R_X(\tau_1)$.

The analytical expression for the autocorrelation function $R_X(\tau)$ shown in Figure 1.6d, is [1]

$$R_X(\tau) = \begin{cases} 1 - \dfrac{|\tau|}{T} & \text{for } |\tau| \leq T \\ 0 & \text{for } |\tau| > T \end{cases} \tag{1.37}$$

Notice that the autocorrelation function gives us frequency information; it tells us something about the bandwidth of the signal. Autocorrelation is a time-domain function; there are no frequency-related terms in the relationship shown in Equation (1.37). How does it give us bandwidth information about the signal? Consider that the signal is a very slowly moving (low bandwidth) signal. As we step the copy along the τ axis, at each step asking the question, "How good is the match between the original and the copy?" the match will be quite good for a while. In other words, the triangular-shaped autocorrelation function in Figure 1.6d and Equation (1.37) will ramp down gradually with τ. But if we have a very rapidly moving (high bandwidth) signal, perhaps a very small shift in τ will result in there being zero correlation. In this case, the autocorrelation function will have a very steep appearance. Therefore, the relative shape of the autocorrelation function tells us something about the bandwidth of the underlying signal. Does it ramp down gently? If so, then we are dealing with a low bandwidth signal. Is the function steep? If so, then we are dealing with a high bandwidth signal.

The autocorrelation function allows us to express a random signal's power spectral density directly. Since the PSD and the autocorrelation function are Fourier transforms of each other, the PSD, $G_X(f)$, of the random bipolar-pulse sequence can be found using Table A.1 as the transform of $R_X(\tau)$ in Equation (1.37). Observe that

$$G_X(f) = T\left(\frac{\sin \pi fT}{\pi fT}\right)^2 = T \operatorname{sinc}^2 fT \tag{1.38}$$

where

$$\operatorname{sinc} y = \frac{\sin \pi y}{\pi y} \tag{1.39}$$

The general shape of $G_X(f)$ is illustrated in Figure 1.6e.

Notice that the area under the PSD curve represents the average power in the signal. One convenient measure of *bandwidth* is the width of the main spectral lobe. (See Section 1.7.2.) Figure 1.6e illustrates that the bandwidth of a signal is inversely related to the symbol duration or pulse width, Figures 1.6f–j repeat the steps shown in Figures 1.6a–e, except that the pulse duration is shorter. Notice that the shape of the shorter pulse duration $R_X(\tau)$ is narrower, shown in Figure 1.6i, than it is for the longer pulse duration $R_X(\tau)$, shown in Figure 1.6d. In Figure 1.6i, $R_X(\tau_1) = 0$; in other words, a shift of τ_1 in the case of the shorter pulse duration example is enough to produce a zero match, or a complete decorrelation between the shifted sequences. Since the pulse duration T is shorter (pulse rate is higher) in Figure 1.6f, than in Figure 1.6a, the bandwidth occupancy in Figure 1.6j is greater than the bandwidth occupancy shown in Figure 1.6e for the lower pulse rate.

1.5.5 Noise in Communication Systems

The term *noise* refers to *unwanted* electrical signals that are always present in electrical systems. The presence of noise superimposed on a signal tends to obscure or mask the signal; it limits the receiver's ability to make correct symbol decisions, and thereby limits the rate of information transmission. Noise arises from a variety of sources, both man made and natural. The *man-made noise* includes such sources as spark-plug ignition noise, switching transients, and other radiating electromagnetic signals. *Natural noise* includes such elements as the atmosphere, the sun, and other galactic sources.

Good engineering design can eliminate much of the noise or its undesirable effect through filtering, shielding, the choice of modulation, and the selection of an optimum receiver site. For example, sensitive radio astronomy measurements are typically located at remote desert locations, far from man-made noise sources. However, there is one natural source of noise, called *thermal* or *Johnson noise,* that cannot be eliminated. Thermal noise [4, 5] is caused by the thermal motion of electrons in all dissipative components—resistors, wires, and so on. The same electrons that are responsible for electrical conduction are also responsible for thermal noise.

We can describe thermal noise as a zero-mean *Gaussian* random process. A Gaussian process $n(t)$ is a random function whose value n at any arbitrary time t is statistically characterized by the Gaussian probability density function

$$p(n) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{n}{\sigma}\right)^2\right] \tag{1.40}$$

where σ^2 is the variance of n. The *normalized* or *standardized Gaussian density function* of a zero-mean process is obtained by assuming that $\sigma = 1$. This normalized pdf is shown sketched in Figure 1.7.

We will often represent a random signal as the sum of a Gaussian noise random variable and a dc signal. That is,

$$z = a + n$$

where z is the random signal, a is the dc component, and n is the Gaussian noise random variable. The pdf $p(z)$ is then expressed as

$$p(z) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{z-a}{\sigma}\right)^2\right] \tag{1.41}$$

where, as before, σ^2 is the variance of n. The Gaussian distribution is often used as the system noise model because of a theorem, called the *central limit theorem* [3], which states that under very general conditions the probability distribution of the sum of j statistically independent random variables approaches the Gaussian distribution as $j \to \infty$, no matter what the individual distribution functions may be. Therefore, even though individual noise mechanisms might have other than Gaussian distributions, the aggregate of many such mechanisms will tend toward the Gaussian distribution.

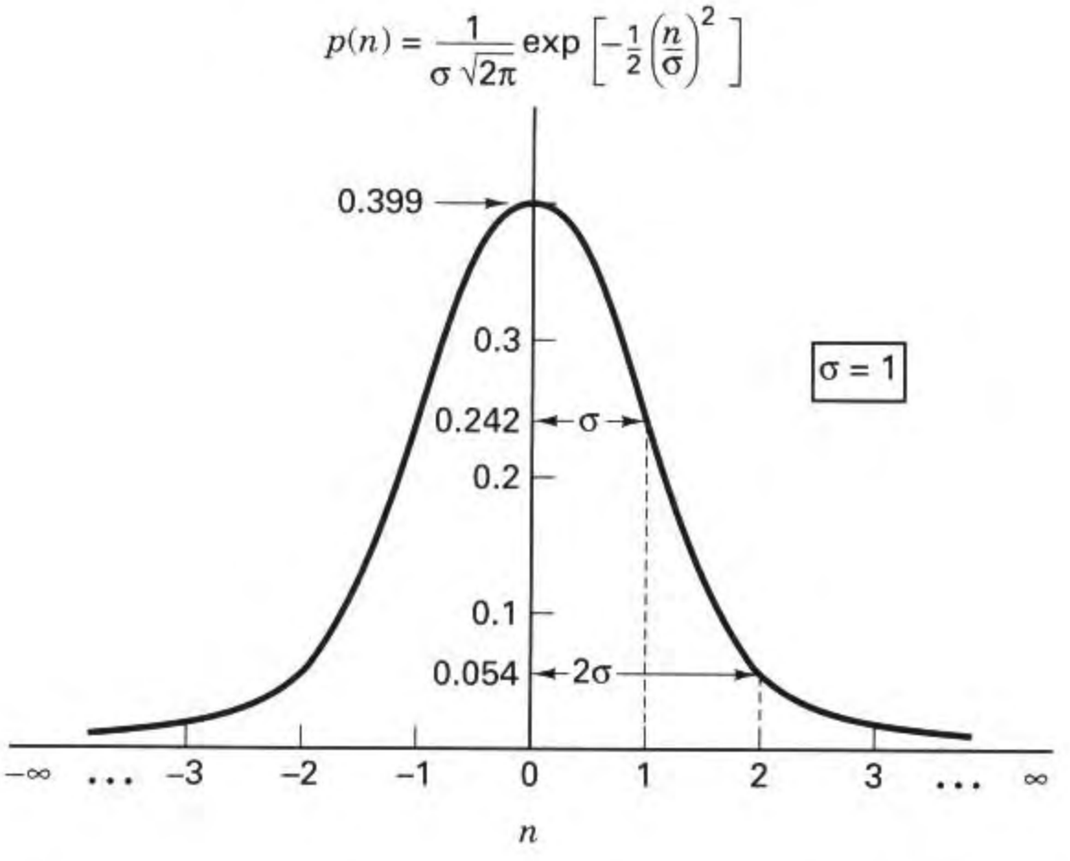


Figure 1.7 Normalized ($\sigma = 1$) Gaussian probability density function.

1.5.5.1 White Noise

The primary spectral characteristic of thermal noise is that its power spectral density is *the same* for all frequencies of interest in most communication systems; in other words, a thermal noise source emanates an equal amount of noise power per unit bandwidth at all frequencies—from dc to about 10^{12} Hz. Therefore, a simple model for thermal noise assumes that its power spectral density $G_n(f)$ is flat for all frequencies, as shown in Figure 1.8a, and is denoted as

$$G_n(f) = \frac{N_0}{2} \qquad \text{watts/hertz} \tag{1.42}$$

where the factor of 2 is included to indicate that $G_n(f)$ is a *two-sided* power spectral density. When the noise power has such a uniform spectral density we refer to it as *white noise*. The adjective "white" is used in the same sense as it is with white light, which contains equal amounts of all frequencies within the visible band of electromagnetic radiation.

The autocorrelation function of white noise is given by the inverse Fourier transform of the noise power spectral density (see Table A.1), denoted as follows:

$$R_n(\tau) = \mathcal{F}^{-1}\{G_n(f)\} = \frac{N_0}{2}\,\delta(\tau) \tag{1.43}$$

Thus the autocorrelation of white noise is a delta function weighted by the factor $N_0/2$ and occurring at $\tau = 0$, as seen in Figure 1.8b. Note that $R_n(\tau)$ is zero for $\tau \neq 0$; that is, any two different samples of white noise, no matter how close together in time they are taken, are uncorrelated.

The average power P_n of white noise is *infinite* because its bandwidth is infinite. This can be seen by combining Equations (1.19) and (1.42) to yield

$$P_n = \int_{-\infty}^{\infty} \frac{N_0}{2}\,df = \infty \tag{1.44}$$

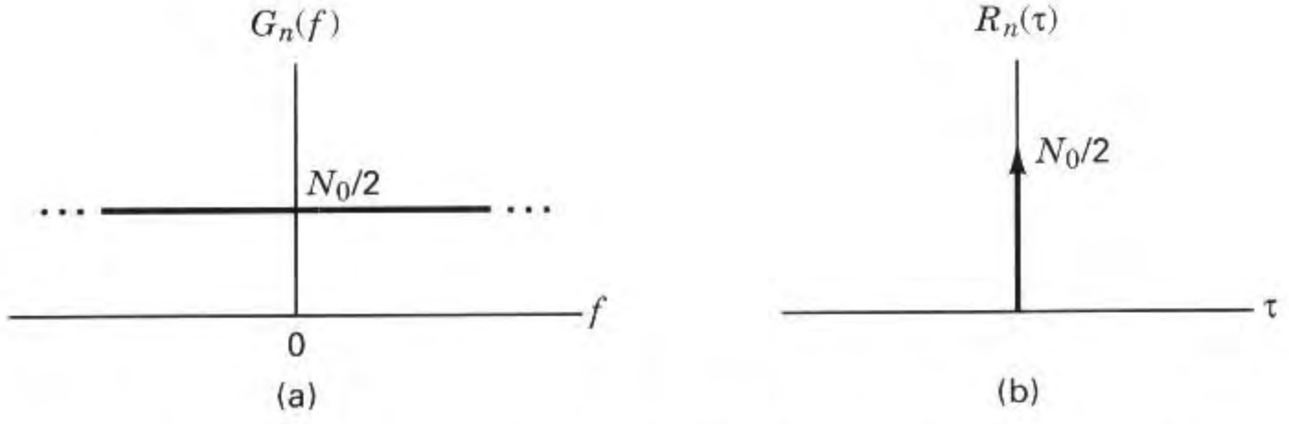


Figure 1.8 (a) Power spectral density of white noise. (b) Autocorrelation function of white noise.

Although white noise is a useful abstraction, no noise process can truly be white; however, the noise encountered in many real systems can be assumed to be approximately white. We can only observe such noise after it has passed through a real system which will have a finite bandwidth. Thus, as long as the bandwidth of the noise is appreciably larger than that of the system, the noise can be considered to have an infinite bandwidth.

The delta function in Equation (1.43) means that the noise signal $n(t)$ is totally decorrelated from its time-shifted version, for any $\tau > 0$. Equation (1.43) indicates that *any* two different samples of a white noise process are uncorrelated. Since thermal noise is a Gaussian process and the samples are uncorrelated, the noise samples are also independent [3]. Therefore, the effect on the detection process of a channel with *additive white Gaussian noise* (AWGN) is that the noise affects each transmitted symbol *independently.* Such a channel is called a *memoryless channel.* The term "additive" means that the noise is simply superimposed or added to the signal—that there are no multiplicative mechanisms at work.

Since thermal noise is present in all communication systems and is the prominent noise source for most systems, the thermal noise characteristics—additive, white, and Gaussian—are most often used to model the noise in communication systems. Since zero-mean Gaussian noise is completely characterized by its *variance,* this model is particularly simple to use in the detection of signals and in the design of optimum receivers. In this book we shall assume, unless otherwise stated, that the system is corrupted by *additive zero-mean white Gaussian noise,* even though this is sometimes an oversimplification.

1.6 SIGNAL TRANSMISSION THROUGH LINEAR SYSTEMS

Having developed a set of models for signals and noise, we now consider the characterization of systems and their effects on such signals and noise. Since a system can be characterized equally well in the time domain or the frequency domain, techniques will be developed in both domains to analyze the response of a linear system to an arbitrary input signal. The signal, applied to the input of the system, as shown in Figure 1.9, can be described either as a time-domain signal, $x(t)$, or by its Fourier transform, $X(f)$. The use of time-domain analysis yields the time-domain output $y(t)$, and in the process, $h(t)$, the characteristic or *impulse response* of the network will be defined. When the input is considered in the frequency domain, we shall define a *frequency transfer function* $H(f)$ for the system, which will determine the frequency-domain output $Y(f)$. The system is assumed to be linear and time invariant. It is also assumed that there is no stored energy in the system at the time the input is applied.

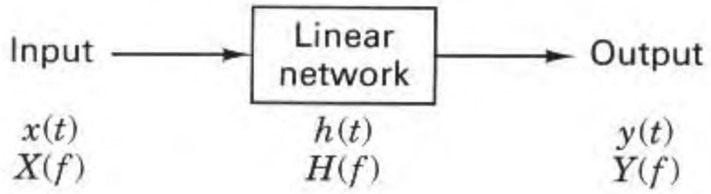


Figure 1.9 Linear system and its key parameters.

1.6.1 Impulse Response

The linear time invariant system or network illustrated in Figure 1.9 is characterized in the time domain by an impulse response $h(t)$, which is the response when the input is equal to a unit impulse $\delta(t)$; that is,

$$h(t) = y(t) \qquad \text{when } x(t) = \delta(t) \tag{1.45}$$

Consider the name *impulse response.* That is a very appropriate name for this event. Characterizing a linear system in terms of its impulse response has a straightforward physical interpretation. At the system input, we apply a unit impulse (a nonrealizable signal, having infinite amplitude, zero width, and unit area), as illustrated in Figure 1.10a. Applying such an impulse to the system can be thought of as giving the system "a whack." How does the system respond to such a force (impulse) at the input? The output response $h(t)$ is the system's impulse response. (A possible shape is depicted in Figure 1.10b.)

The response of the network to an arbitrary input signal $x(t)$ is found by the convolution of $x(t)$ with $h(t)$, expressed as

$$y(t) = x(t) * h(t) = \int_{-\infty}^{\infty} x(\tau)\, h(t - \tau)\, d\tau \tag{1.46}$$

where $*$ denotes the convolution operation. (See Section A.5.) The system is assumed to be *causal,* which means that there can be *no* output prior to the time, $t = 0$, when the input is applied. Therefore, the lower limit of integration can be changed to zero, and we can express the output $y(t)$ in either the form

$$y(t) = \int_{0}^{\infty} x(\tau) h(t - \tau)\, d\tau \tag{1.47a}$$

or the form

$$y(t) = \int_{0}^{\infty} x(t - \tau)\, h(\tau)\, d\tau \tag{1.47b}$$

Each of the expressions in Equations (1.46) and (1.47) is called the *convolution integral.* Convolution is a basic mathematical tool that plays an important role in understanding all communication systems. Thus, the reader is urged to review

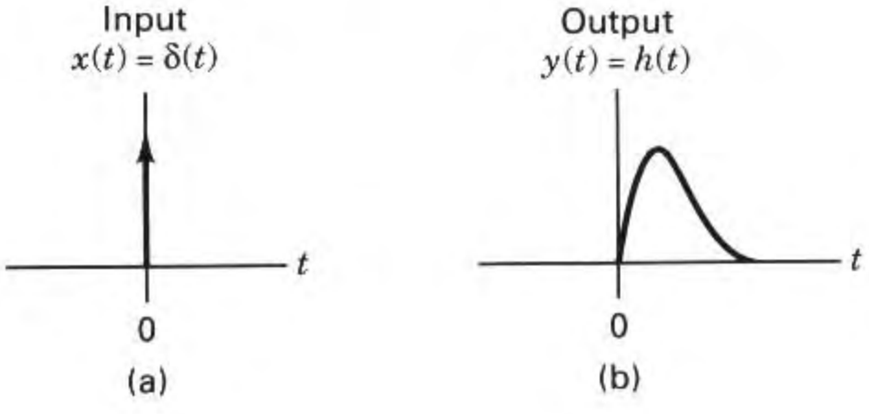


Figure 1.10 (a) Input signal $x(t)$ is a unit impulse function. (b) Output signal $y(t)$ is the system's impulse response $h(t)$.

section A.5, where one can see that Equations (1.46) and (1.47) are the results of a straightforward process.

1.6.2 Frequency Transfer Function

The frequency-domain output signal $Y(f)$ is obtained by taking the Fourier transform of both sides of Equation (1.46). Since convolution in the time domain transforms to multiplication in the frequency domain (and vice versa), Equation (1.46) yields

$$Y(f) = X(f)H(f) \tag{1.48}$$

or

$$H(f) = \frac{Y(f)}{X(f)} \tag{1.49}$$

provided, of course, that $X(f) \neq 0$ for all f. Here $H(f) = \mathcal{F}\{h(t)\}$, the Fourier transform of the impulse response function, is called the *frequency transfer function* or the *frequency response* of the network. In general, $H(f)$ is complex and can be written as

$$H(f) = |H(f)|\, e^{j\theta(f)} \tag{1.50}$$

where $|H(f)|$ is the magnitude response. The phase response is defined as

$$\theta(f) = \tan^{-1} \frac{\text{Im}\,\{H(f)\}}{\text{Re}\,\{H(f)\}} \tag{1.51}$$

where the terms "Re" and "Im" denote "the real part of" and "the imaginary part of," respectively.

The frequency transfer function of a linear time-invariant network can easily be measured in the laboratory with a sinusoidal generator at the input of the network and an oscilloscope at the output. When the input waveform $x(t)$ is expressed as

$$x(t) = A \cos 2\pi f_0 t$$

the output of the network will be

$$y(t) = A\,|H(f_0)| \cos\,[2\pi f_0 t + \theta(f_0)] \tag{1.52}$$

The input frequency f_0 is stepped through the values of interest; at each step, the amplitude and phase at the output are measured.

1.6.2.1 Random Processes and Linear Systems

If a random process forms the input to a time-invariant linear system, the output will also be a random process. That is, each sample function of the input process yields a sample function of the output process. The input power spectral density $G_X(f)$ and the output power spectral density $G_Y(f)$ are related as follows:

$$G_Y(f) = G_X(f)\,|H(f)|^2 \tag{1.53}$$

Equation (1.53) provides a simple way of finding the power spectral density out of a time-invariant linear system when the input is a random process.

In Chapters 3 and 4 we consider the detection of signals in Gaussian noise. We will utilize a fundamental property of a Gaussian process applied to a linear system, as follows. It can be shown that if a Gaussian process $X(t)$ is applied to a time-invariant linear filter, the random process $Y(t)$ developed at the output of the filter is also Gaussian [6].

1.6.3 Distortionless Transmission

What is required of a network for it to behave like an *ideal* transmission line? The output signal from an ideal transmission line may have some time delay compared with the input, and it may have a different amplitude than the input (just a scale change), but otherwise it must have no distortion—it must have the same shape as the input. Therefore, for ideal distortionless transmission, we can describe the output signal as

$$y(t) = Kx(t - t_0) \tag{1.54}$$

where K and t_0 are constants. Taking the Fourier transform of both sides (see Section A.3.1), we write

$$Y(f) = KX(f)e^{-j2\pi f t_0} \tag{1.55}$$

Substituting the expression (1.55) for $Y(f)$ into Equation (1.49), we see that the required system transfer function for distortionless transmission is

$$H(f) = Ke^{-j2\pi f t_0} \tag{1.56}$$

Therefore, to achieve *ideal distortionless transmission,* the overall system response must have a constant magnitude response and its phase shift must be linear with frequency. It is not enough that the system amplify or attenuate all frequency components equally. All of the signal's frequency components must also arrive with identical time delay in order to add up correctly. Since the time delay t_0 is related to the phase shift θ and the radian frequency $\omega = 2\pi f$ by

$$t_0\,(\text{seconds}) = \frac{\theta\,(\text{radians})}{2\pi f\,(\text{radians/second})} \tag{1.57a}$$

it is clear that phase shift must be proportional to frequency in order for the time delay of all components to be identical. A characteristic often used to measure delay distortion of a signal is called *envelope delay* or *group delay,* which is defined as

$$\tau(f) = -\frac{1}{2\pi}\frac{d\theta(f)}{df} \tag{1.57b}$$

Therefore, for distortionless transmission, an equivalent way of characterizing phase to be a linear function of frequency is to characterize the envelope delay $\tau(f)$ as a constant. In practice, a signal will be distorted in passing through some parts of a system. Phase or amplitude correction (*equalization*) networks may be introduced elsewhere in the system to correct for this distortion. It is the overall input–output characteristic of the system that determines its performance.

1.6.3.1 Ideal Filter

One cannot build the ideal network described in Equation (1.56). The problem is that Equation (1.56) implies an infinite bandwidth capability, where the bandwidth of a system is defined as the interval of positive frequencies over which the magnitude $|H(f)|$ remains within a specified value. In Section 1.7 various measures of bandwidth are enumerated. As an approximation to the ideal infinite-bandwidth network, let us choose a truncated network that passes, without distortion, all frequency components between f_ℓ and f_u, where f_ℓ is the lower cutoff frequency and f_u is the upper cutoff frequency, as shown in Figure 1.11. Each of these networks is called an *ideal filter.* Outside the range $f_\ell < f < f_u$, which is called the *passband,* the ideal filter is assumed to have a response of zero magnitude. The effective width of the passband is specified by the filter bandwidth $W_f = (f_u - f_\ell)$ hertz.

When $f_\ell \neq 0$ and $f_u \neq \infty$, the filter is called a *bandpass filter* (BPF), shown in Figure 1.11a. When $f_\ell = 0$ and f_u has a finite value, the filter is called a *low-pass filter* (LPF), shown in Figure 1.11b. When f_ℓ has a nonzero value and when $f_u \to \infty$, the filter is called a *high-pass filter* (HPF), shown in Figure 1.11c.

Following Equation (1.56) and letting $K = 1$, for the ideal low-pass filter transfer function with bandwidth $W_f = f_u$ hertz, shown in Figure 1.11b, we can write the transfer function as

$$H(f) = |H(f)|\, e^{-j\theta(f)} \tag{1.58}$$

where

$$|H(f)| = \begin{cases} 1 & \text{for } |f| < f_u \\ 0 & \text{for } |f| \geq f_u \end{cases} \tag{1.59}$$

and

$$e^{-j\,\theta(f)} = e^{-j\,2\pi f t_0} \tag{1.60}$$

The impulse response of the ideal low-pass filter, illustrated in Figure 1.12, is

$$h(t) = \mathcal{F}^{-1}\{H(f)\} = \int_{-\infty}^{\infty} H(f) e^{j\,2\pi f t}\, df \tag{1.61}$$

$$= \int_{-f_u}^{f_u} e^{-j\,2\pi f t_0} e^{j\,2\pi f t}\, df$$

or

$$= \int_{-f_u}^{f_u} e^{j2\pi f(t-t_0)}\, df$$

$$= 2f_u \frac{\sin 2\pi f_u(t - t_0)}{2\pi f_u(t - t_0)}$$

$$= 2f_u \operatorname{sinc} 2f_u(t - t_0) \tag{1.62}$$

where sinc x is as defined in Equation (1.39). The impulse response shown in Figure 1.12 is noncausal, which means that it has a nonzero output prior to the

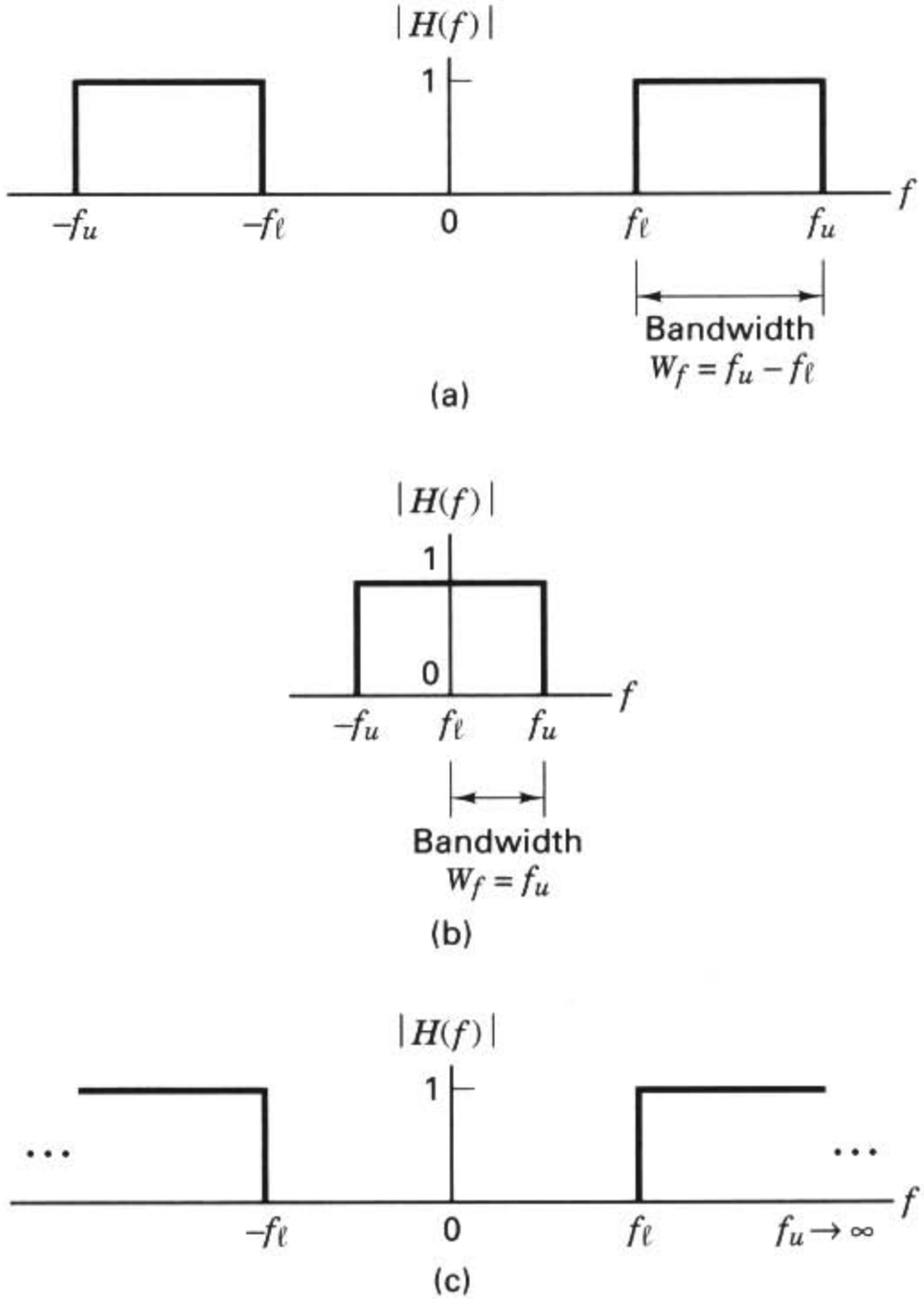


Figure 1.11 Ideal filter transfer function. (a) Ideal bandpass filter. (b) Ideal low-pass filter. (c) Ideal high-pass filter.

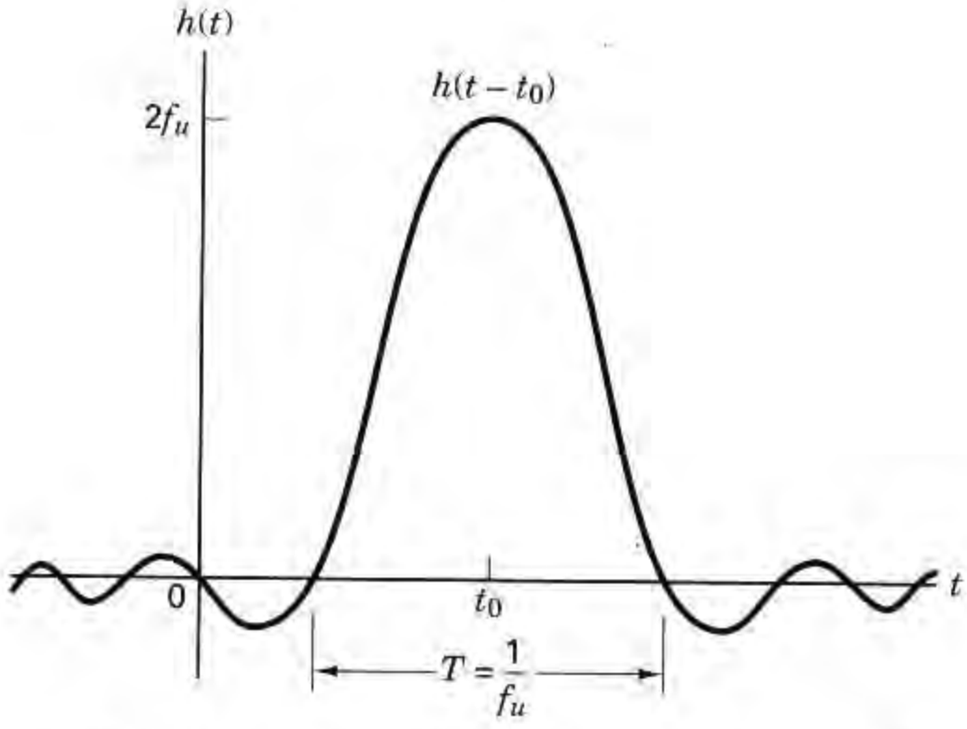


Figure 1.12 Impulse response of the ideal low-pass filter.

application of an input at time $t = 0$. Therefore, it should be clear that the ideal filter described in Equation (1.58) is not realizable.

Example 1.2 Effect of an Ideal Filter on White Noise

White noise with power spectral density $G_n(f) = N_0/2$, shown in Figure 1.8a, forms the input to the ideal low-pass filter shown in Figure 1.11b. Find the power spectral density $G_Y(f)$ and the autocorrelation function $R_Y(\tau)$ of the output signal.

Solution

$$G_Y(f) = G_n(f)\,|H(f)|^2$$

$$= \begin{cases} \dfrac{N_0}{2} & \text{for } |f| < f_u \\ 0 & \text{otherwise} \end{cases}$$

The autocorrelation is the inverse Fourier transform of the power spectral density and is given by (see Table A.1)

$$R_Y(\tau) = N_0 f_u \frac{\sin 2\pi f_u \tau}{2\pi f_u \tau}$$

$$= N_0 f_u \operatorname{sinc} 2f_u \tau$$

Comparing this result with Equation (1.62), we see that $R_Y(\tau)$ has the same shape as the impulse response of the ideal low-pass filter shown in Figure 1.12. In this example the ideal low-pass filter transforms the autocorrelation function of white noise (defined by the delta function) into a sinc function. After filtering, we no longer have white noise. The output noise signal will have zero correlation with shifted copies of itself, only at shifts of $\tau = n/2f_u$, where n is any integer other than zero.

1.6.3.2 Realizable Filters

The very simplest example of a realizable low-pass filter is made up of resistance ($\mathcal{R}$) and capacitance (C), as shown in Figure 1.13a; it is called an $\mathcal{R}C$ *filter*, and its transfer function can be expressed as [7]

$$H(f) = \frac{1}{1 + j2\pi f\mathcal{R}C} = \frac{1}{\sqrt{1 + (2\pi f\mathcal{R}C)^2}}\, e^{-j\theta(f)} \tag{1.63}$$

where $\theta(f) = \tan^{-1} 2\pi f\mathcal{R}C$. The magnitude characteristic $|H(f)|$ and the phase characteristic $\theta(f)$ are plotted in Figures 1.13b and c, respectively. The low-pass filter bandwidth is defined to be its half-power point; this point is the frequency at which the output signal power has fallen to one-half of its peak value, or the frequency at which the magnitude of the output voltage has fallen to $1/\sqrt{2}$ of its peak value.

The half-power point is generally expressed in decibel (dB) units as the –3-dB point, or the point that is 3 dB down from the peak, where the decibel is defined as the ratio of two amounts of power, P_1 and P_2, existing at two points. By definition,

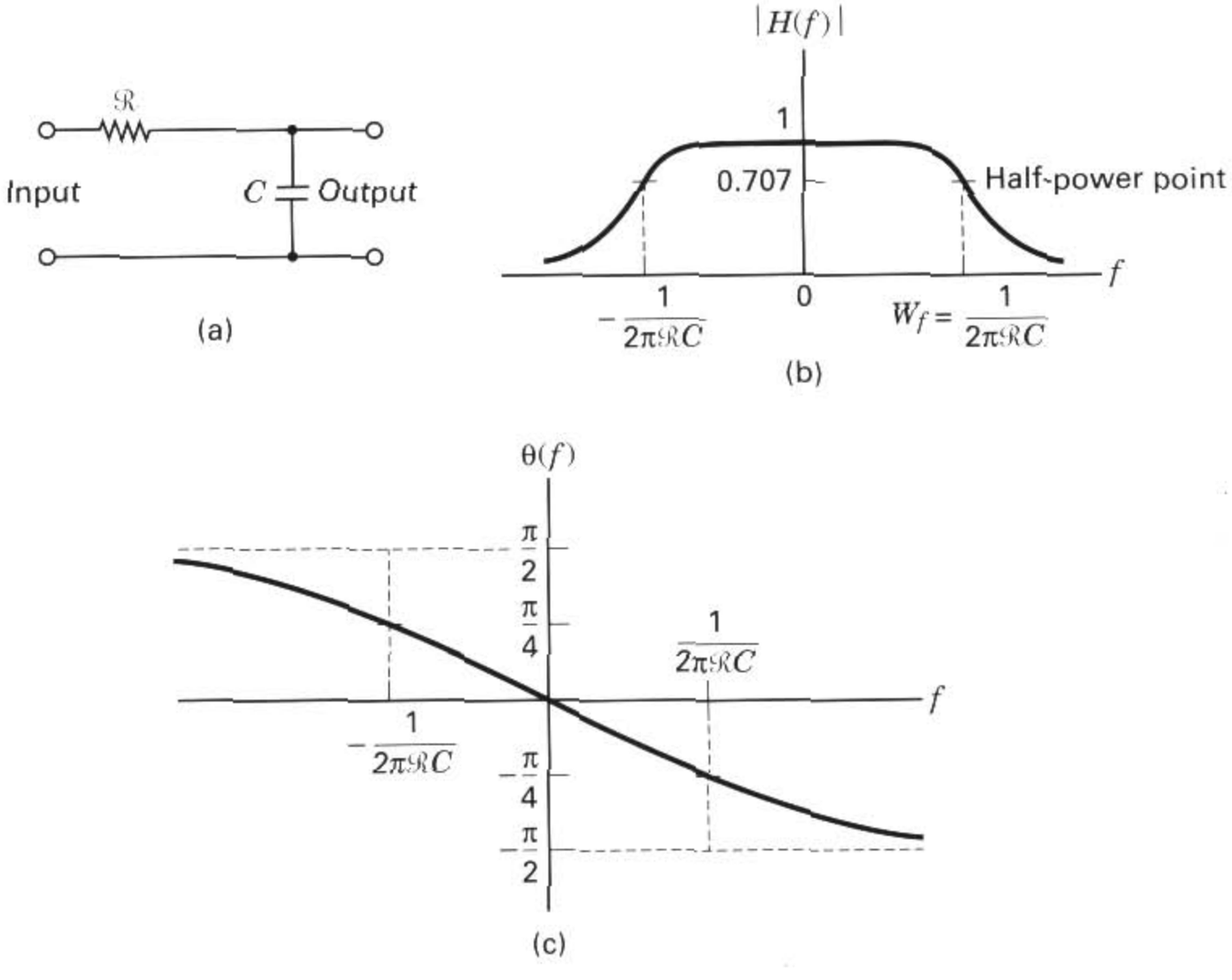


Figure 1.13 $\mathcal{R}C$ filter and its transfer function. (a) $\mathcal{R}C$ filter. (b) Magnitude characteristic of the $\mathcal{R}C$ filter. (c) Phase characteristic of the $\mathcal{R}C$ filter.

$$\text{number of dB} = 10 \log_{10} \frac{P_2}{P_1} = 10 \log_{10} \frac{V_2^2/\mathcal{R}_2}{V_1^2/\mathcal{R}_1} \tag{1.64a}$$

where V_1 and V_2 are voltages and $\mathcal{R}_1$ and $\mathcal{R}_2$ are resistances. For communication systems, *normalized power* is generally used for analysis; in this case, $\mathcal{R}_1$ and $\mathcal{R}_2$ are set equal to 1 Ω, so that

$$\text{number of dB} = 10 \log_{10} \frac{P_2}{P_1} = 10 \log_{10} \frac{V_2^2}{V_1^2} \tag{1.64b}$$

The amplitude response can be expressed in decibels by

$$|H(f)|_{\text{dB}} = 20 \log_{10} \frac{V_2}{V_1} = 20 \log_{10} |H(f)| \tag{1.64c}$$

where V_1 and V_2 are the input and output voltages, respectively, and where the input and output resistances have been assumed equal.

From Equation (1.63) it is easy to verify that the half-power point of the low-pass $\mathcal{R}C$ filter corresponds to $\omega = 1/\mathcal{R}C$ radians per second or $f = 1/(2\pi\mathcal{R}C)$ hertz. Thus the bandwidth W_f in hertz is $1/(2\pi\mathcal{R}C)$. The filter *shape factor* is a measure of how well a realizable filter approximates the ideal filter. It is typically defined as the ratio of the filter bandwidths at the –60-dB and –6-dB amplitude response points. A sharp-cutoff bandpass filter can be made with a shape factor as low as about 2. By comparison, the shape factor of the simple $\mathcal{R}C$ low-pass filter is almost 600.

There are several useful approximations to the ideal low-pass filter characteristic. One of these, the *Butterworth filter,* approximates the ideal low-pass filter with the function

$$|H_n(f)| = \frac{1}{\sqrt{1 + (f/f_u)^{2n}}} \qquad n \geq 1 \tag{1.65}$$

where f_u is the upper –3-db cutoff frequency and n is referred to as the *order* of the filter. The higher the order, the greater will be the complexity and the cost to implement the filter. The magnitude function, $|H(f)|$, is sketched (single sided) for several values of n in Figure 1.14. Note that as n gets larger, the magnitude characteristics approach that of the ideal filter. Butterworth filters are popular because they are the best approximation to the ideal, in the sense of *maximal flatness* in the filter passband.

Example 1.3 Effect of an $\mathcal{R}C$ Filter on White Noise

White noise with spectral density $G_n(f) = N_0/2$, shown in Figure 1.8a, forms the input to the $\mathcal{R}C$ filter shown in Figure 1.13a. Find the power spectral density $G_Y(f)$ and the autocorrelation function $R_Y(\tau)$ of the output signal.

Solution

$$\begin{aligned} G_Y(f) &= G_n(f)\,|H(f)|^2 \\ &= \frac{N_0}{2} \frac{1}{1 + (2\pi f \mathcal{R}C)^2} \\ R_Y(\tau) &= \mathcal{F}^{-1}\{G_Y(f)\} \end{aligned}$$

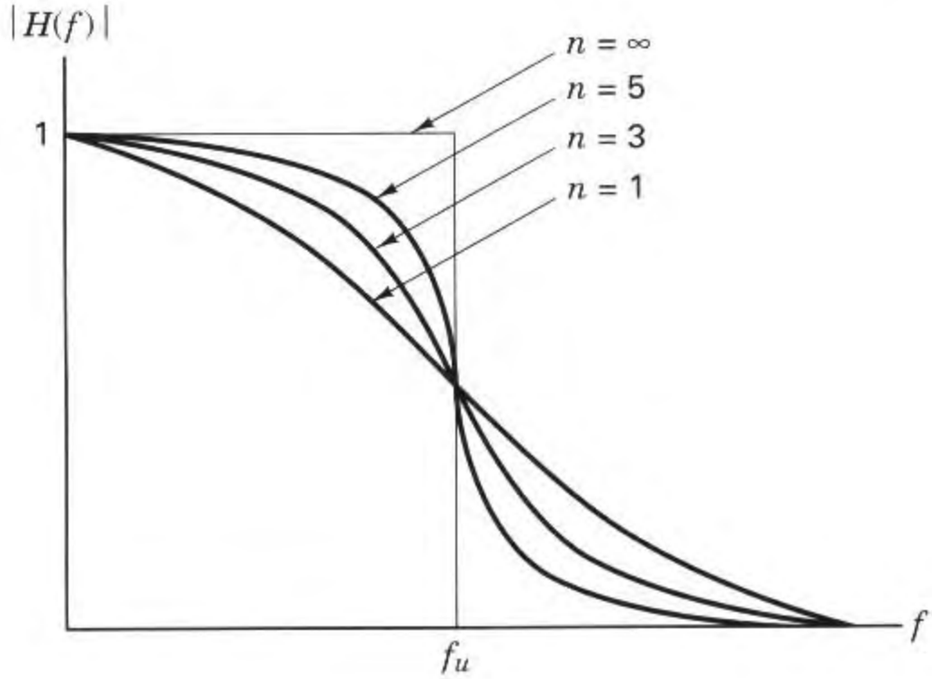


Figure 1.14 Butterworth filter magnitude response.

Using Table A.1, we find that the inverse Fourier transform of $G_Y(f)$ is

$$R_Y(\tau) = \frac{N_0}{4\mathcal{R}C} \exp\left(-\frac{|\tau|}{\mathcal{R}C}\right)$$

As might have been predicted, we no longer have white noise after filtering. The $\mathcal{R}C$ filter transforms the input autocorrelation function of white noise (defined by the delta function) into an exponential function. For a narrowband filter (a large $\mathcal{R}C$ product), the output noise will exhibit higher correlation between noise samples of a fixed time shift than will the output noise from a wideband filter.

1.6.4 Signals, Circuits, and Spectra

Signals have been described in terms of their spectra. Similarly, networks or circuits have been described in terms of their spectral characteristics or frequency transfer functions. How is a signal's bandwidth affected as a result of the signal passing through a filter circuit? Figure 1.15 illustrates two cases of interest. In Figure 1.15a (case 1), the input signal has a narrowband spectrum, and the filter transfer function is a wideband function. From Equation (1.48), we see that the output signal spectrum is simply the product of these two spectra. In Figure 1.15a we can verify that multiplication of the two spectral functions will result in a spectrum with a bandwidth approximately equal to the smaller of the two bandwidths (when one of the two spectral functions goes to zero, the multiplication yields zero). Therefore, for case 1, the output signal spectrum is constrained by the input signal spectrum alone. Similarly, we see that for case 2, in Figure 1.15b, where the input signal is a wideband signal but the filter has a narrowband transfer function, the bandwidth of the output signal is constrained by the filter bandwidth; the output signal will be a filtered (distorted) rendition of the input signal.

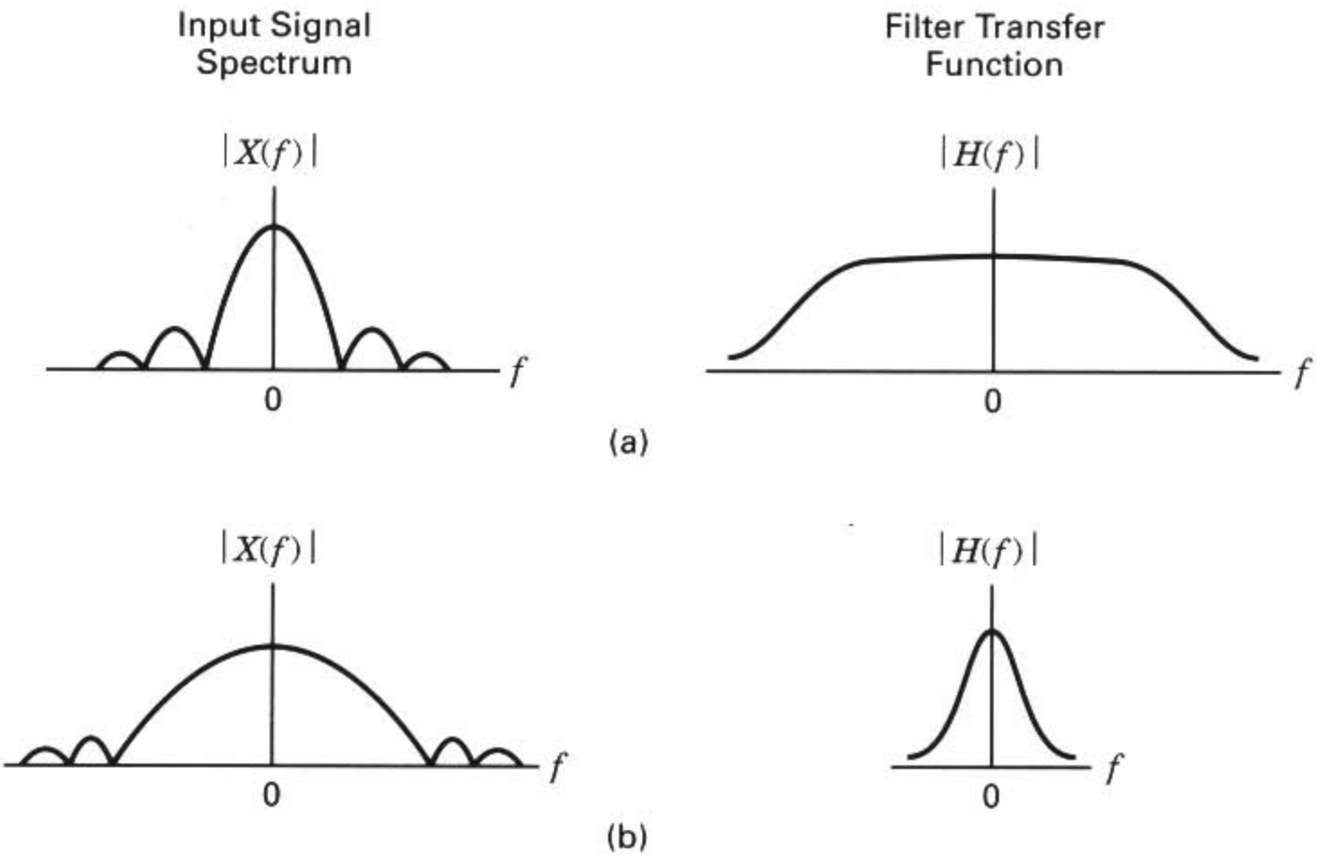


Figure 1.15 Spectral characteristics of the input signal and the circuit contribute to the spectral characteristics of the output signal. (a) Case 1: Output bandwidth is constrained by input signal bandwidth. (b) Case 2: Output bandwidth is constrained by filter bandwidth.

The effect of a filter on a waveform can also be viewed in the time domain. The output $y(t)$ resulting from convolving an ideal input pulse $x(t)$ (having amplitude V_m and pulse width T) with the impulse response of a low-pass $\mathcal{R}C$ filter can be written as [8]

$$y(t) = \begin{cases} V_m(1 - e^{-t/\mathcal{R}C}) & \text{for } 0 \leq t \leq T \\ V'_m e^{-(t-T)/\mathcal{R}C} & \text{for } t > T \end{cases} \tag{1.66}$$

where

$$V'_m = V_m(1 - e^{-T/\mathcal{R}C}) \tag{1.67}$$

Let us define the pulse bandwidth as

$$W_p = \frac{1}{T} \tag{1.68}$$

and the $\mathcal{R}C$ filter bandwidth, as

$$W_f = \frac{1}{2\pi\mathcal{R}C} \tag{1.69}$$

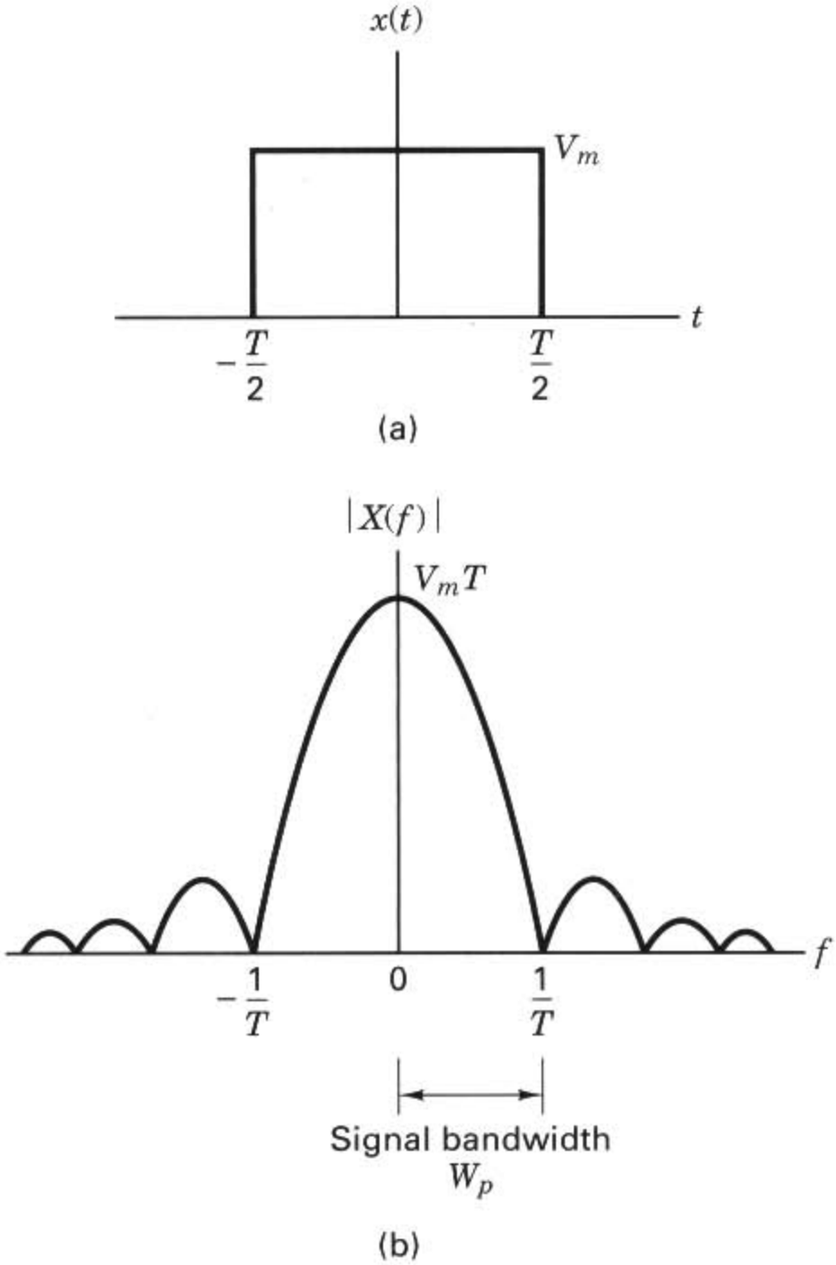


Figure 1.16 (a) Ideal pulse. (b) Magnitude spectrum of the ideal pulse.

The ideal input pulse $x(t)$ and its magnitude spectrum $|X(f)|$ are shown in Figure 1.16. The $\mathcal{R}C$ filter and its magnitude characteristic $|H(f)|$ are shown in Figures 1.13a and b, respectively. Following Equations (1.66) to (1.69), three cases are illustrated in Figure 1.17. Example 1 illustrates the case where $W_p << W_f$. Notice that the output response $y(t)$ is a reasonably good approximation of the input pulse $x(t)$, shown in dashed lines. This represents an example of *good fidelity.* In example 2, where $W_p \approx W_f$, we can still recognize that a pulse had been transmitted from the output $y(t)$. Finally, example 3 illustrates the case in which $W_p >> W_f$. Here the presence of the pulse is barely perceptible from $y(t)$. Can you think of an application where the large filter bandwidth or good fidelity of example 1 is called for? A *precise ranging application,* perhaps, where the pulse time of arrival translates into distance, necessitates a pulse with a steep rise time. Which example characterizes the binary digital communications application? *It is example 2.* As we pointed out earlier regarding Figure 1.1, one of the principal features of binary digital communications is that each received pulse need only be accurately *perceived* as being in one of its two states; a high-fidelity signal need not be maintained. Example 3 has

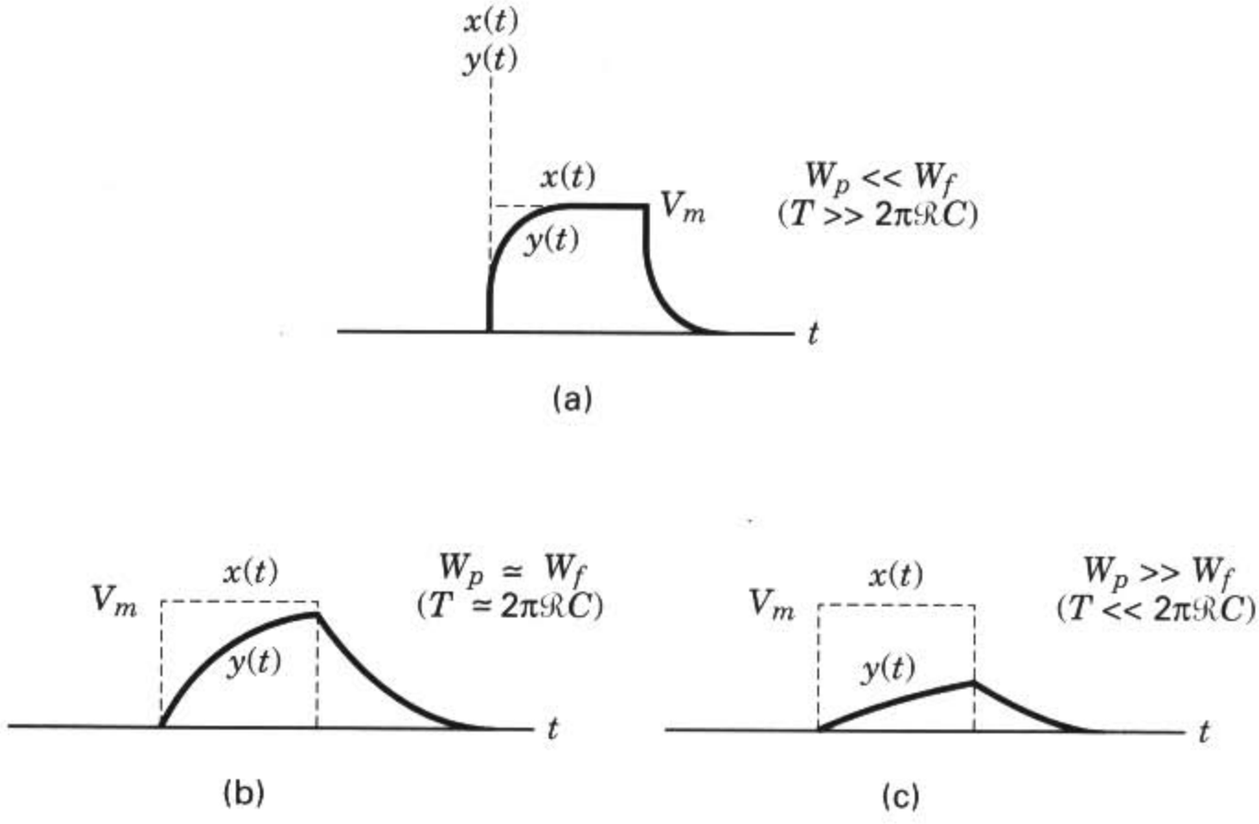


Figure 1.17 Three examples of filtering an ideal pulse. (a) Example 1: Good-fidelity output. (b) Example 2: Good-recognition output. (c) Example 3: Poor-recognition output.

been included for completeness; it would not be used as a design criterion for a practical system.

.7 BANDWIDTH OF DIGITAL DATA

1.7.1 Baseband versus Bandpass

An easy way to translate the spectrum of a low-pass or baseband signal $x(t)$ to a higher frequency is to multiply or *heterodyne* the baseband signal with a carrier wave $\cos 2\pi f_c t$, as shown in Figure 1.18. The resulting waveform, $x_c(t)$, is called a *double-sideband* (DSB) *modulated signal* and is expressed as

$$x_c(t) = x(t) \cos 2\pi f_c t \tag{1.70}$$

From the frequency shifting theorem (see Section A.3.2), the spectrum of the DSB signal $x_c(t)$ is given by

$$X_c(f) = \tfrac{1}{2}[X(f - f_c) + X(f + f_c)] \tag{1.71}$$

The magnitude spectrum $|X(f)|$ of the baseband signal $x(t)$ having a bandwidth f_m and the magnitude spectrum $|X_c(f)|$ of the DSB signal $x_c(t)$ having a bandwidth W_{DSB} are shown in Figure 1.18b and c, respectively. In the plot of $|X_c(f)|$, spectral

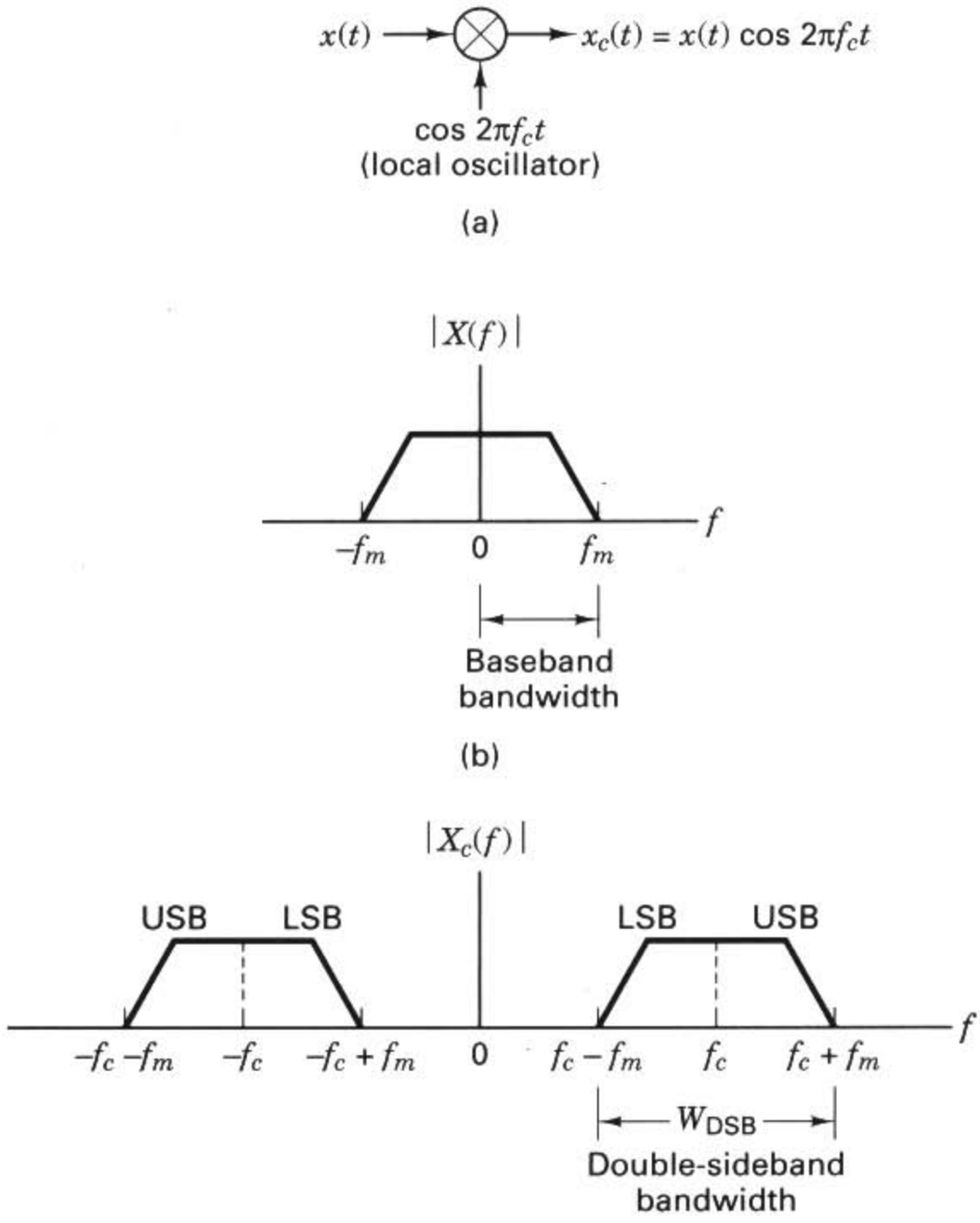


Figure 1.18 Comparison of baseband and double-sideband spectra. (a) Heterodyning. (b) Baseband spectrum. (c) Double-sideband spectrum.

components corresponding to positive baseband frequencies appear in the range f_c to $(f_c + f_m)$. This part of the DSB spectrum is called the *upper sideband* (USB). Spectral components corresponding to negative baseband frequencies appear in the range $(f_c - f_m)$ to f_c. This part of the DSB spectrum is called the *lower sideband* (LSB). Mirror images of the USB and LSB spectra appear in the negative-frequency half of the plot. The *carrier wave* is sometimes referred to as a *local oscillator* (LO) *signal,* a *mixing signal,* or a *heterodyne signal.* Generally, the carrier wave frequency is much higher than the bandwidth of the baseband signal; that is,

$$f_c \gg f_m$$

From Figure 1.18, we can readily compare the bandwidth f_m required to transmit the baseband signal with the bandwidth W_{DSB} required to transmit the DSB signal; we see that

$$W_{\text{DSB}} = 2f_m \tag{1.72}$$

That is, we need twice as much transmission bandwidth to transmit a DSB version of the signal than we do to transmit its baseband counterpart.

1.7.2 The Bandwidth Dilemma

Many important theorems of communication and information theory are based on the assumption of *strictly bandlimited* channels, which means that no signal power whatever is allowed outside the defined band. We are faced with the dilemma that strictly bandlimited signals, as depicted by the spectrum $|X_1(f)|$ in Figure 1.19b, are not realizable, because they imply signals with infinite duration, as seen by $x_1(t)$ in Figure 1.19a (the inverse Fourier transform of $X_1(f)$. Duration-limited signals, as seen by $x_2(t)$ in Figure 1.19c, can clearly be realized. However, such signals are just as unreasonable, since their Fourier transforms contain energy at arbitrarily high frequencies as depicted by the spectrum $|X_2(f)|$ in Figure 1.19d. In summary, for all bandlimited spectra, the waveforms are not realizable, and for all realizable waveforms, the absolute bandwidth is infinite. The mathematical description of a real signal does not permit the signal to be strictly duration limited and strictly bandlimited. Hence, the mathematical models are abstractions; it is no wonder that there is no single universal definition of bandwidth.

All bandwidth criteria have in common the attempt to specify a measure of the width, W, of a nonnegative real-valued spectral density defined for all frequen-

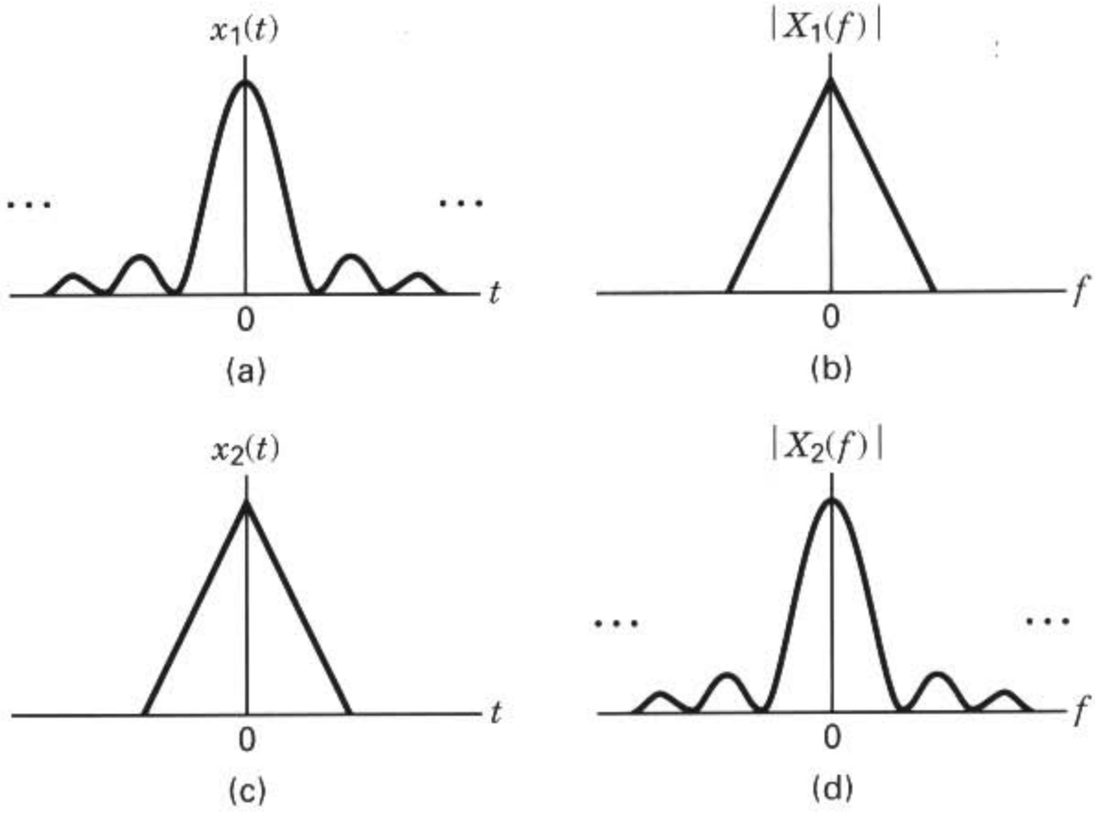


Figure 1.19 (a) Strictly bandlimited signal in the time domain. (b) In the frequency domain. (c) Strictly time limited signal in the time domain. (d) In the frequency domain.

cies $|f| < \infty$. Figure 1.20 illustrates some of the most common definitions of bandwidth; in general, the various criteria are not interchangeable. The single-sided power spectral density for a single heterodyned pulse $x_c(t)$ takes the analytical form

$$G_x(f) = T\left[\frac{\sin \pi(f - f_c)T}{\pi(f - f_c)T}\right]^2 \tag{1.73}$$

where f_c is the carrier wave frequency and T is the pulse duration. This power spectral density, whose general appearance is sketched in Figure 1.20, also characterizes a *random pulse sequence,* assuming that the averaging time is long relative to the pulse duration. The plot consists of a main lobe and smaller symmetrical sidelobes. The general shape of the plot is valid for most digital modulation formats; some formats, however, do not have well-defined lobes. The bandwidth criteria depicted in Figure 1.20 are as follows:

(a) *Half-power bandwidth.* This is the interval between frequencies at which $G_x(f)$ has dropped to half-power, or 3 dB below the peak value.

(b) *Equivalent rectangular or noise equivalent bandwidth.* The noise equivalent bandwidth was originally conceived to permit rapid computation of output noise power from an amplifier with a wideband noise input; the concept can similarly be applied to a signal bandwidth. The noise equivalent bandwidth

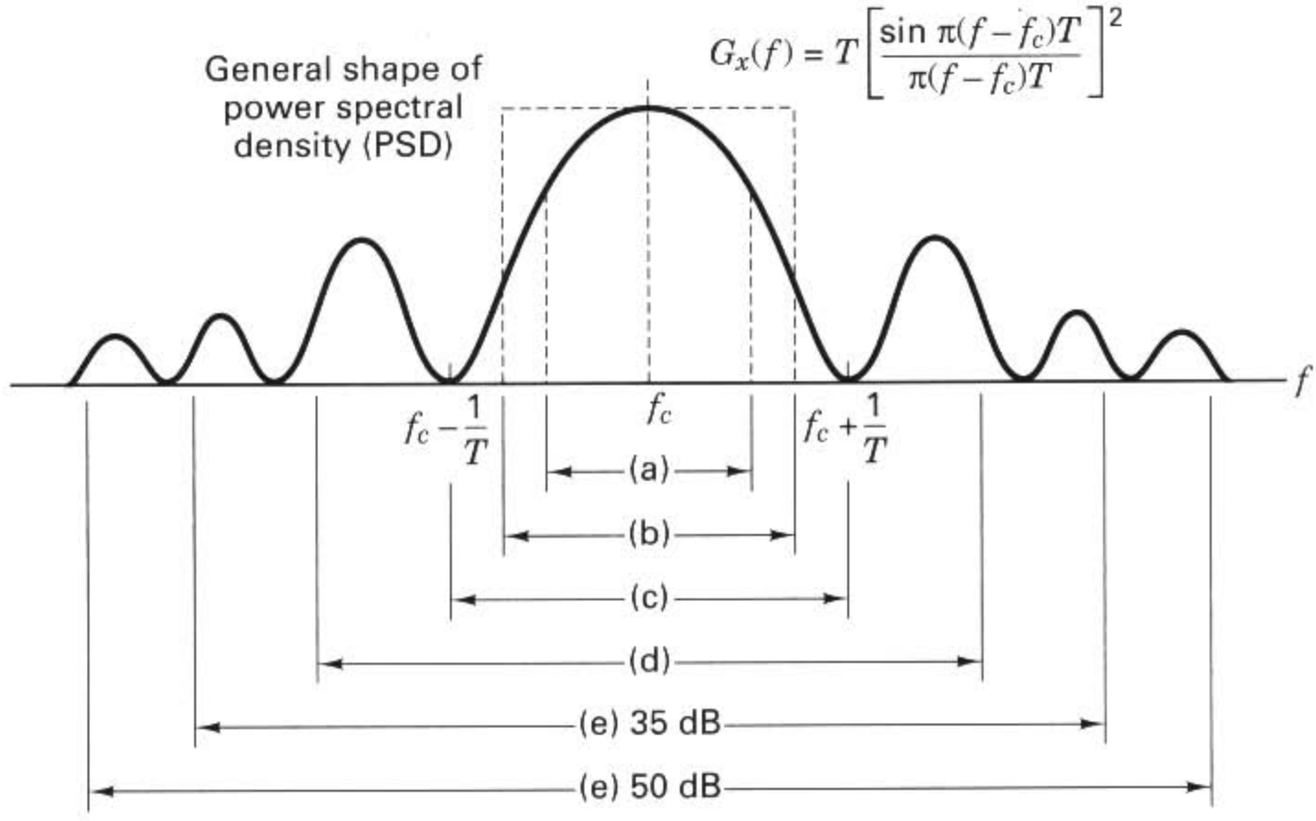


Figure 1.20 Bandwidth of digital data. (a) Half-power. (b) Noise equivalent. (c) Null to null. (d) 99% of power. (e) Bounded PSD (defines attenuation outside bandwidth) at 35 and 50 dB.

W_N of a signal is defined by the relationship $W_N = P_x/G_x(f_c)$, where P_x is the total signal power over all frequencies and $G_x(f_c)$ is the value of $G_x(f)$ at the band center (assumed to be the maximum value over all frequencies).

(c) *Null-to-null bandwidth.* The most popular measure of bandwidth for digital communications is the width of the main spectral lobe, where most of the signal power is contained. This criterion lacks complete generality since some modulation formats lack well-defined lobes.

(d) *Fractional power containment bandwidth.* This bandwidth criterion has been adopted by the Federal Communications Commission (FCC Rules and Regulations Section 2.202) and states that the occupied bandwidth is the band that leaves exactly 0.5% of the signal power above the upper band limit and exactly 0.5% of the signal power below the lower band limit. Thus 99% of the signal power is inside the occupied band.

(e) *Bounded power spectral density.* A popular method of specifying bandwidth is to state that everywhere outside the specified band, $G_x(f)$ must have fallen at least to a certain stated level below that found at the band center. Typical attenuation levels might be 35 or 50 dB.

(f) *Absolute bandwidth.* This is the interval between frequencies, outside of which the spectrum is zero. This is a useful abstraction. However, for all realizable waveforms, the absolute bandwidth is infinite.

Example 1.4 Strictly Bandlimited Signals

The concept of a signal that is strictly limited to a band of frequencies is not realizable. Prove this by showing that a *strictly bandlimited* signal must also be a signal of *infinite time duration.*

Solution

Let $x(t)$ be a signal with Fourier transform $X(f)$ that is strictly limited to the band of frequencies centered at $\pm f_c$ and of width $2W$. We may express $X(f)$ in terms of an ideal filter transfer function $H(f)$, illustrated in Figure 1.21a, as

$$X(f) = X'(f)H(f) \tag{1.74}$$

where $X'(f)$ is the Fourier transform of a signal $x'(t)$, not necessarily bandlimited, and

$$H(f) = \text{rect}\left(\frac{f - f_c}{2W}\right) + \text{rect}\left(\frac{f + f_c}{2W}\right) \tag{1.75}$$

in which

$$\text{rect}\left(\frac{f}{2W}\right) = \begin{cases} 1 & \text{for } -W < f < W \\ 0 & \text{for } |f| > W \end{cases}$$

We can express $X(f)$ in terms of $X'(f)$ as

$$X(f) = \begin{cases} X'(f) & \text{for } (f_c - W) \le |f_c| \le (f_c + W) \\ 0 & \text{otherwise} \end{cases}$$

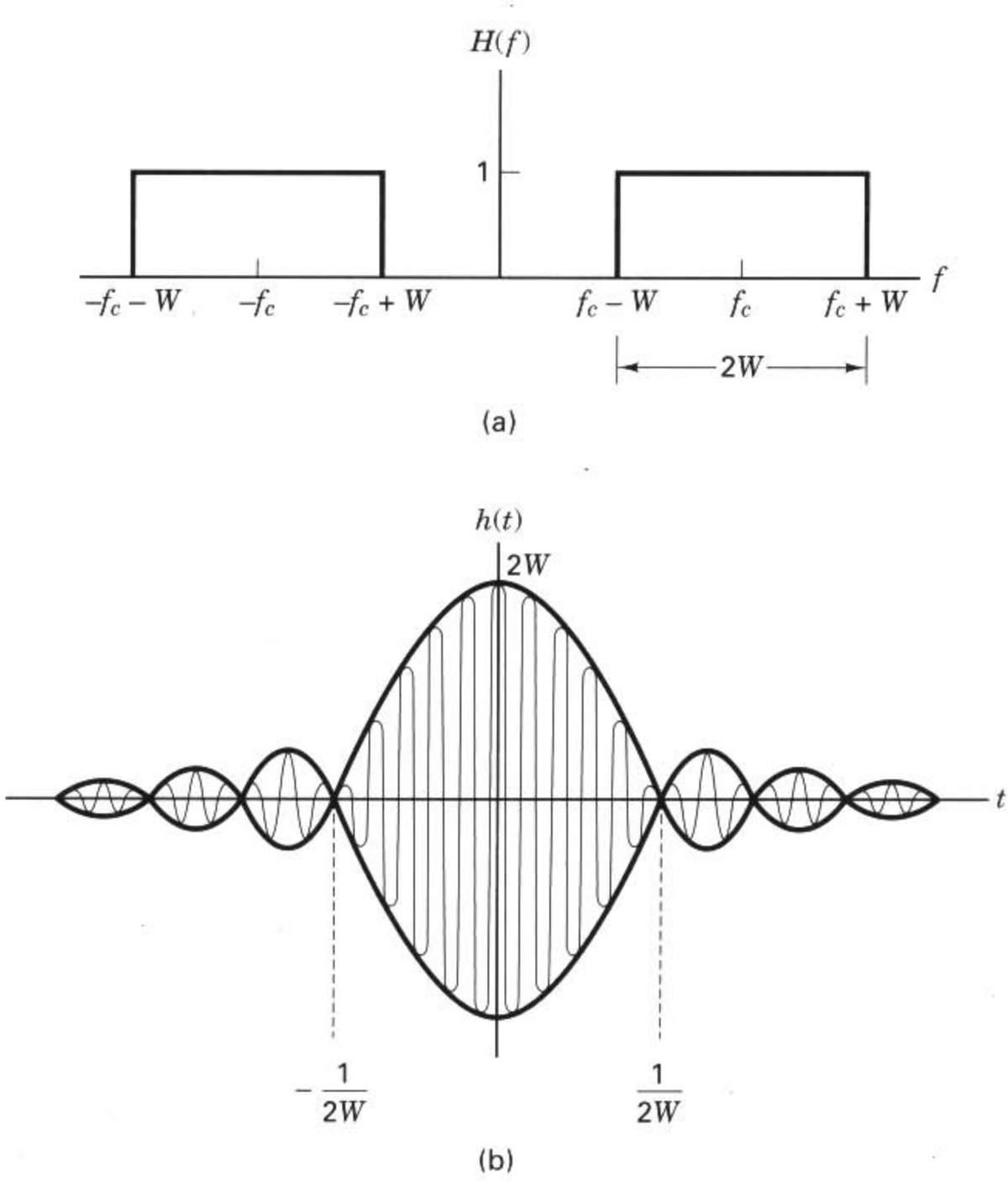


Figure 1.21 Transfer function and impulse response for a strictly bandlimited signal. (a) Ideal bandpass filter. (b) Ideal bandpass impulse response.

Multiplication in the frequency domain, as seen in Equation (1.74), transforms to convolution in the time domain as

$$x(t) = x'(t) * h(t) \tag{1.76}$$

where $h(t)$, the inverse Fourier transform of $H(f)$, can be written as (see Tables A.1 and A.2)

$$h(t) = 2W\,(\text{sinc}\,2Wt)\cos 2\pi f_c t$$

and is illustrated in Figure 1.21b. We note that $h(t)$ is of *infinite duration.* It follows, therefore, that $x(t)$ obtained in Equation (1.76) by convolving $x'(t)$ with $h(t)$ is also of infinite duration and therefore is *not realizable.*

1.8 CONCLUSION

In this chapter, the goals of the book have been outlined and the basic nomenclature has been defined. The fundamental concepts of time-varying signals, such as classification, spectral density, and autocorrelation, have been reviewed. Also, random signals have been considered, and white Gaussian noise, the primary noise model in most communication systems, has been characterized, statistically and spectrally. Finally, we have treated the important area of signal transmission through linear systems and have examined some of the realizable approximations to the ideal case. We have also established that the concept of an absolute bandwidth is an abstraction, and that in the real world we are faced with the need to choose a definition of bandwidth that is useful for our particular application. In the remainder of the book, each of the signal processing steps introduced in this chapter will be explored in the context of the typical system block diagram appearing at the beginning of each chapter.

REFERENCES

1. Haykin, S., *Communication Systems,* John Wiley & Sons, Inc., New York, 1983.
2. Shanmugam, K. S., *Digital and Analog Communication Systems,* John Wiley & Sons, Inc., New York, 1979.
3. Papoulis, A., *Probability, Random Variables, and Stochastic Processes,* McGraw-Hill Book Company, New York, 1965.
4. Johnson, J. B., "Thermal Agitation of Electricity in Conductors," *Phys. Rev.,* vol. 32, July 1928, pp. 97–109.
5. Nyguist, H., "Thermal Agitation of Electric Charge in Conductors," *Phys. Rev.,* vol. 32, July 1928, pp. 110–113.
6. Van Trees, H. L., *Detection, Estimation, and Modulation Theory,* Part 1, John Wiley & Sons, New York, 1968.
7. Schwartz, M., *Information Transmission, Modulation, and Noise,* McGraw-Hill Book Company, New York, 1970.
8. Millman, J., and Taub, H., *Pulse, Digital, and Switching Waveforms,* McGraw-Hill Book Company, New York, 1965.

PROBLEMS

1.1. Classify the following signals as energy signals or power signals. Find the normalized energy or normalized power of each.

(a) $x(t) = A \cos 2\pi f_0 t$ for $-\infty < t < \infty$

(b) $x(t) = \begin{cases} A \cos 2\pi f_0 t & \text{for } -T_0/2 \le t \le T_0/2, \text{ where } T_0 = 1/f_0 \\ 0 & \text{elsewhere} \end{cases}$

(c) $x(t) = \begin{cases} A \exp(-at) & \text{for } t > 0, a > 0 \\ 0 & \text{elsewhere} \end{cases}$

(d) $x(t) = \cos t + 5 \cos 2t$ for $-\infty < t < \infty$

1.2. Determine the energy spectral density of a square pulse $x(t) = \text{rect}\ (t/T)$, where rect (t/T) equals 1, for $-T/2 \le t \le T/2$, and equals 0, elsewhere. Calculate the normalized energy E_x in the pulse.

1.3. Find an expression for the average normalized power in a periodic signal in terms of its complex Fourier series coefficients.

1.4. Using time averaging, find the average normalized power in the waveform $x(t) = 10 \cos 10t + 20 \cos 20t$.

1.5. Repeat Problem 1.4 using the summation of spectral coefficients.

1.6. Determine which, if any, of the following functions have the properties of autocorrelation functions. Justify your determination. [*Note:* $\mathcal{F}\{R(\tau)\}$ must be a nonnegative function. Why?]

(a) $x(\tau) = \begin{cases} 1 & \text{for } -1 \le \tau \le 1 \\ 0 & \text{otherwise} \end{cases}$

(b) $x(\tau) = \delta(\tau) + \sin 2\pi f_0 \tau$

(c) $x(\tau) = \exp(|\tau|)$

(d) $x(\tau) = 1 - |\tau|$ for $-1 \le \tau \le 1$, 0 elsewhere

1.7. Determine which, if any, of the following functions have the properties of power spectral density functions. Justify your determination.

(a) $X(f) = \delta(f) + \cos^2 2\pi f$

(b) $X(f) = 10 + \delta(f - 10)$

(c) $X(f) = \exp(-2\pi |f - 10|)$

(d) $X(f) = \exp[-2\pi(f^2 - 10)]$

1.8. Find the autocorrelation function of $x(t) = A \cos(2\pi f_0 t + \phi)$ in terms of its period, $T_0 = 1/f_0$. Find the average normalized power of $x(t)$, using $P_x = R(0)$.

1.9. **(a)** Use the results of Problem 1.8 to find the autocorrelation function $R(\tau)$ of waveform $x(t) = 10 \cos 10t + 20 \cos 20t$.

(b) Use the relationship $P_x\ R(0)$ to find the average normalized power in $x(t)$. Compare the answer with the answers to Problems 1.4 and 1.5.

1.10. For the function $x(t) = 1 + \cos 2\pi f_0 t$, calculate (**a**) the average value of $x(t)$; (**b**) the ac power of $x(t)$; (**c**) the rms value of $x(t)$.

1.11. Consider a random process given by $X(t) = A \cos(2\pi f_0 t + \phi)$, where A and f_0 are constants and ϕ is a random variable that is uniformly distributed over $(0, 2\pi)$. If $X(t)$ is an ergodic process, the time averages of $X(t)$ in the limit as $t \to \infty$ are equal to the corresponding ensemble averages of $X(t)$.

(a) Use time averaging over an integer number of periods to calculate the approximations to the first and second moments of $X(t)$.

(b) Use Equations (1.26) and (1.28) to calculate the ensemble-average approximations to the first and second moments of $X(t)$. Compare the results with your answers in part (a).

1.12. The Fourier transform of a signal $x(t)$ is defined by $X(f) = \text{sinc}\ f$, where the sinc function is as defined in Equation (1.39). Find the autocorrelation function, $R_x(\tau)$, of the signal $x(t)$.

1.13. Use the sampling property of the unit impulse function to evaluate the following integrals.

(a) $\int_{-\infty}^{\infty} \cos 6t\delta(t - 3)\, dt$

(b) $\int_{-\infty}^{\infty} 10\delta(t)(1 + t)^{-1}\, dt$

(c) $\int_{-\infty}^{\infty} \delta(t + 4)(t^2 + 6t + 1)\, dt$

(d) $\int_{-\infty}^{\infty} \exp(-t^2)\delta(t - 2)\, dt$

1.14. Find $X_1(f) * X_2(f)$ for the spectra shown in Figure P1.1.

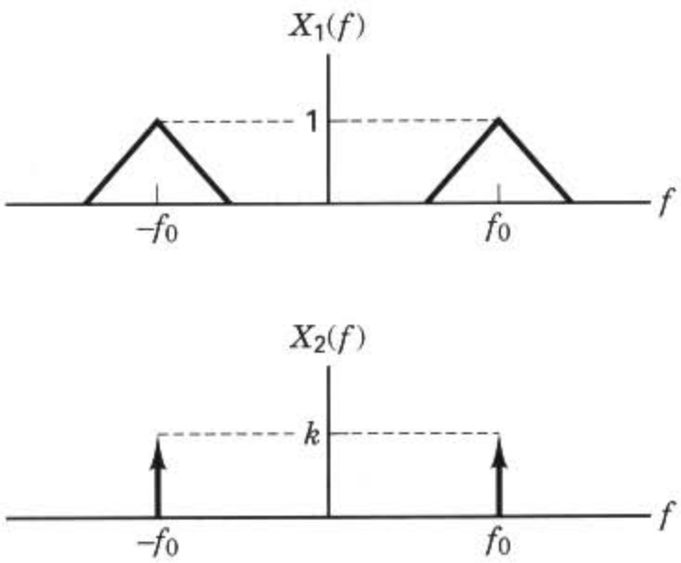


Figure P1.1

1.15. The two-sided power spectral density, $G_x(f) = 10^{-6} f^2$, of a waveform $x(t)$ is shown in Figure P1.2.

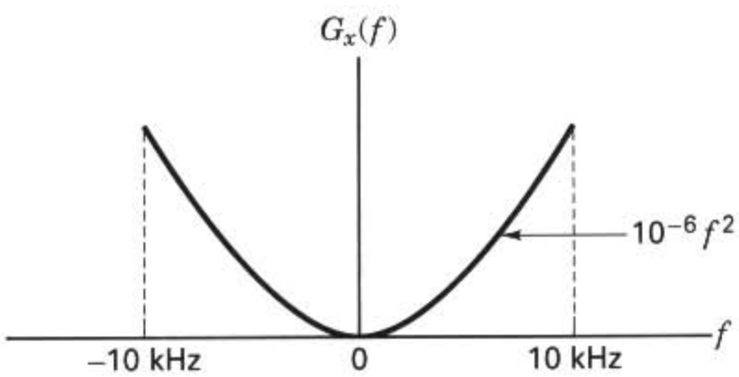


Figure P1.2

(a) Find the normalized average power in $x(t)$ over the frequency band from 0 to 10 kHz.

(b) Find the normalized average power contained in the frequency band from 5 to 6 kHz.

1.16. Decibels are logarithmic measures of *power ratios,* as described in Equation (1.64a). Sometimes, a similar formulation is used to express nonpower measurements in decibels (referenced to some designated unit). As an example, calculate how many decibels of hamburger meat you would buy to feed 2 hamburgers each to a group of 100 people. Assume that you and the butcher have agreed on the unit of "½ pound of meat" (the amount in one hamburger) as a reference unit.

1.17. Consider the Butterworth low-pass amplitude response given in Equation (1.65).

(a) Find the value of n so that $|H(f)|^2$ is constant to within ± 1 dB over the range $|f| \leq 0.9f_u$.

(b) Show that as n approaches infinity, the amplitude response approaches that of an ideal low-pass filter.

1.18. Consider the network in Figure 1.9, whose frequency transfer function is $H(f)$. An impulse $\delta(t)$ is applied at the input. Show that the response $y(t)$ at the output is the inverse Fourier transform of $H(f)$.

1.19. An example of a *holding circuit,* commonly used in pulse systems, is shown in Figure P1.3. Determine the impulse response of this circuit.

1.20. Given the spectrum

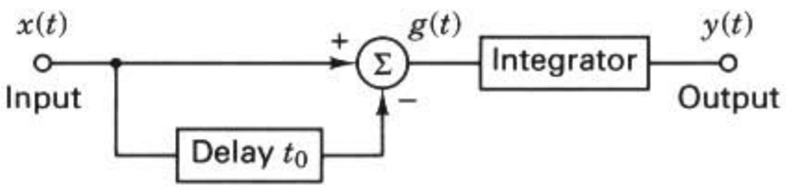


Figure P1.3

$$G_x(f) = 10^{-4}\left\{\frac{\sin\left[\pi(f - 10^6)10^{-4}\right]}{\pi(f - 10^6)10^{-4}}\right\}^2$$

find the value of the signal bandwidth using the following bandwidth definitions:

(a) Half-power bandwidth.

(b) Noise equivalent bandwidth.

(c) Null-to-null bandwidth.

(d) 99% of power bandwidth. (Hint: Use numerical methods.)

(e) Bandwidth beyond which the attenuation is 35 dB.

(f) Absolute bandwidth.

QUESTIONS

1.1. How does the plot of a signal's autocorrelation function reveal its bandwidth occupancy? (See Section 1.5.4.)

1.2. What two requirements must be fulfilled in order to insure distortionless transmission through a linear system? (See Section 1.6.3.)

1.3. Define the parameter *envelope delay* or *group delay.* (See Section 1.6.3.)

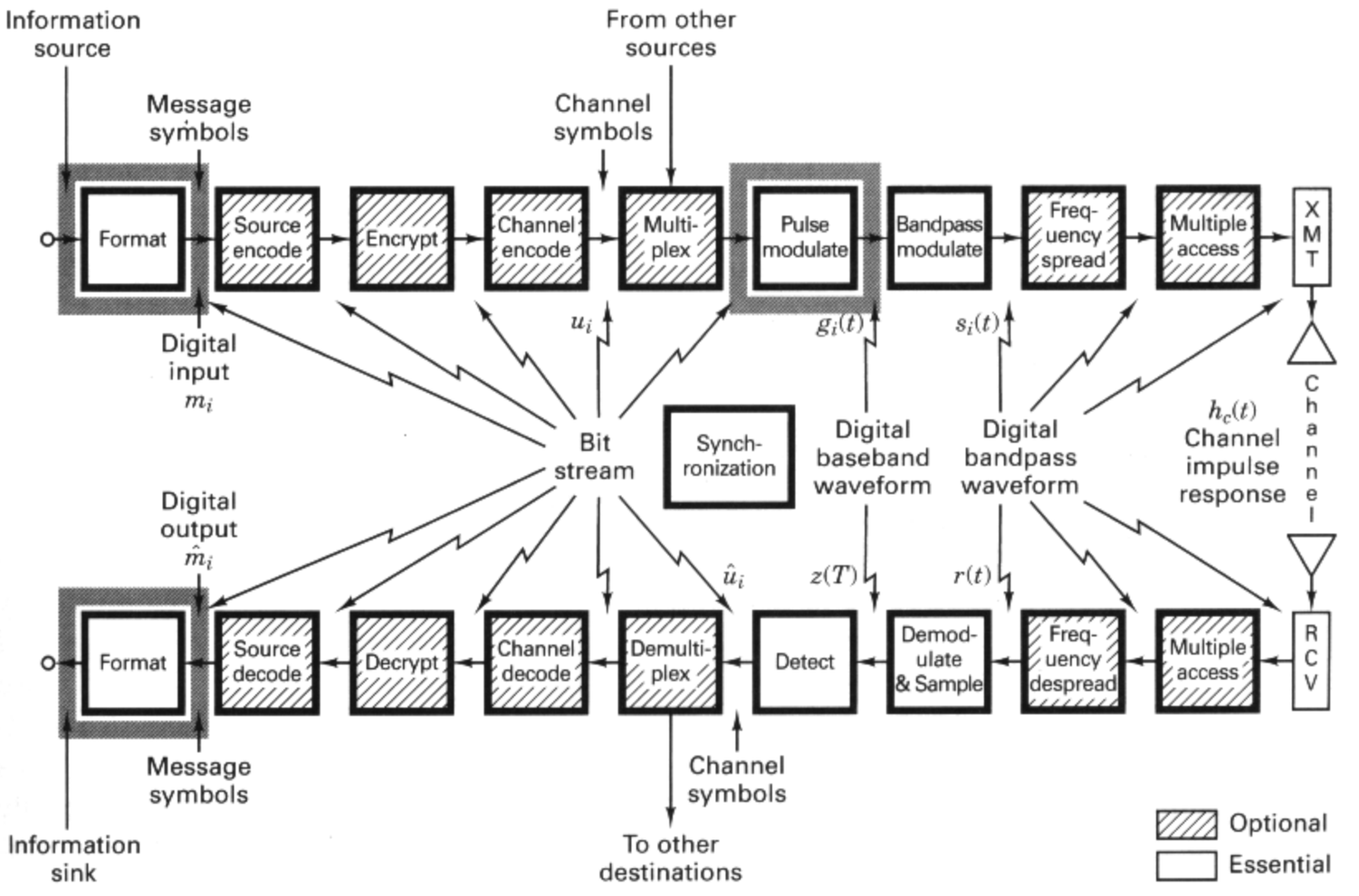
1.4. What mathematical dilemma is the cause for there being several different definitions of bandwidth? (See Section 1.7.2.)

EXERCISES

Using the Companion CD, run the exercises associated with Chapter 1.

2 FORMATTING AND BASEBAND MODULATION 55

Formatting and Baseband Modulation



The goal of the first essential signal processing step, *formatting*, is to insure that the message (or source signal) is compatible with digital processing. *Transmit formatting* is a transformation from source information to digital symbols. (It is the reverse transformation in the receive chain.) When data compression in addition to formatting is employed, the process is termed *source coding*. Some authors consider formatting a special case of source coding. We treat formatting (and baseband modulation) in this chapter, and treat source coding as a special case of the *efficient description* of source information in Chapter 13.

In Figure 2.1, the highlighted block labeled "formatting" contains a list of topics that deal with transforming information to digital messages. The digital messages are considered to be in the logical format of binary ones and zeros until they are transformed by the next essential step, called pulse modulation, into *baseband* (pulse) waveforms. Such waveforms can then be transmitted over a cable. In Figure 2.1, the highlighted block labeled "baseband signaling" contains a list of pulse modulating waveforms that are described in this chapter. The term baseband refers to a signal whose spectrum extends from (or near) dc up to some finite value, usually less than a few megahertz. In Chapter 3, the subject of baseband signaling is continued with emphasis on demodulation and detection.

2.1 BASEBAND SYSTEMS

In Figure 1.2 we presented a block diagram of a typical digital communication system. A version of this functional diagram, focusing primarily on the formatting and transmission of *baseband* signals, is shown in Figure 2.2. Data already in a digital

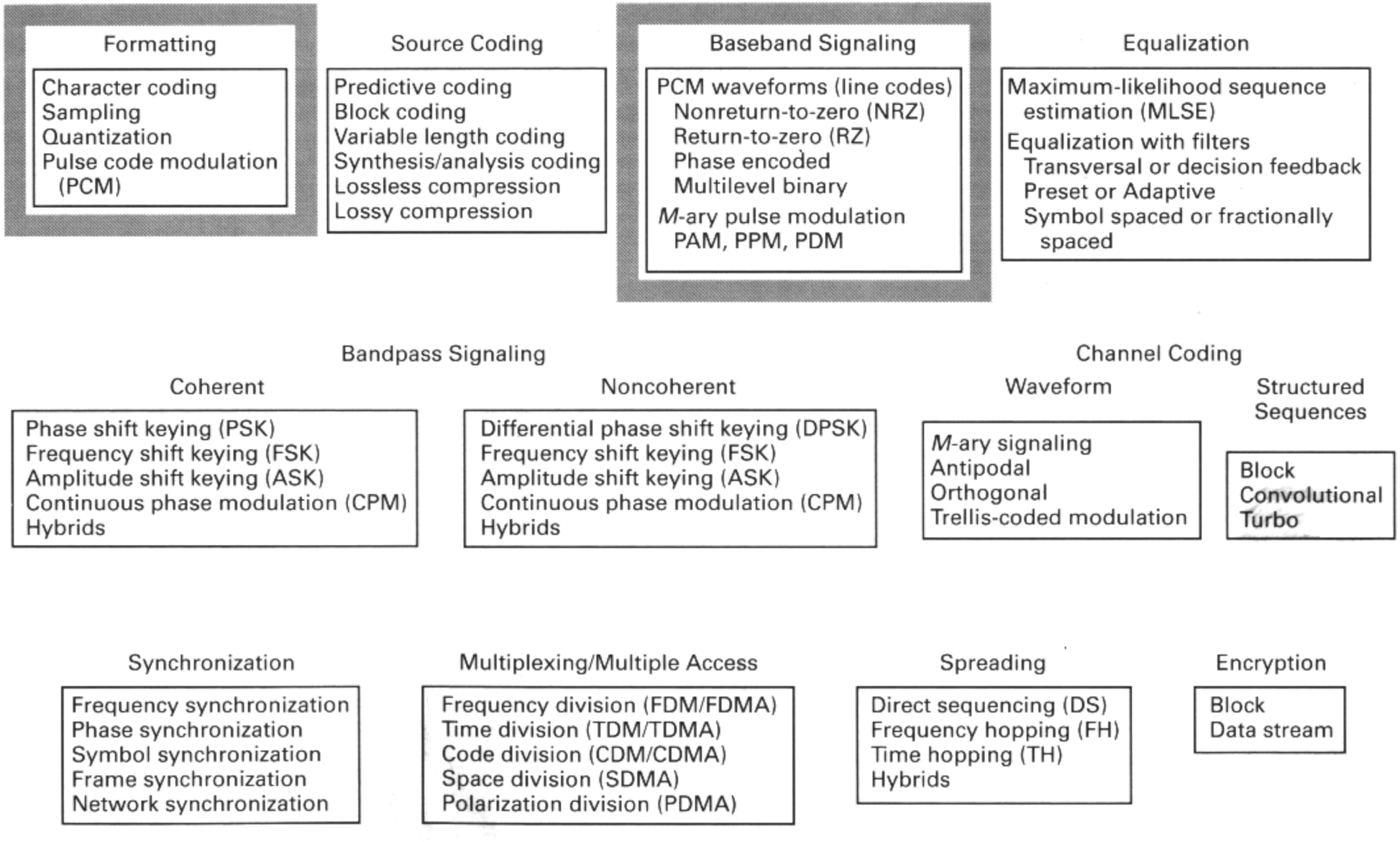


Figure 2.1 Basic digital communication transformations

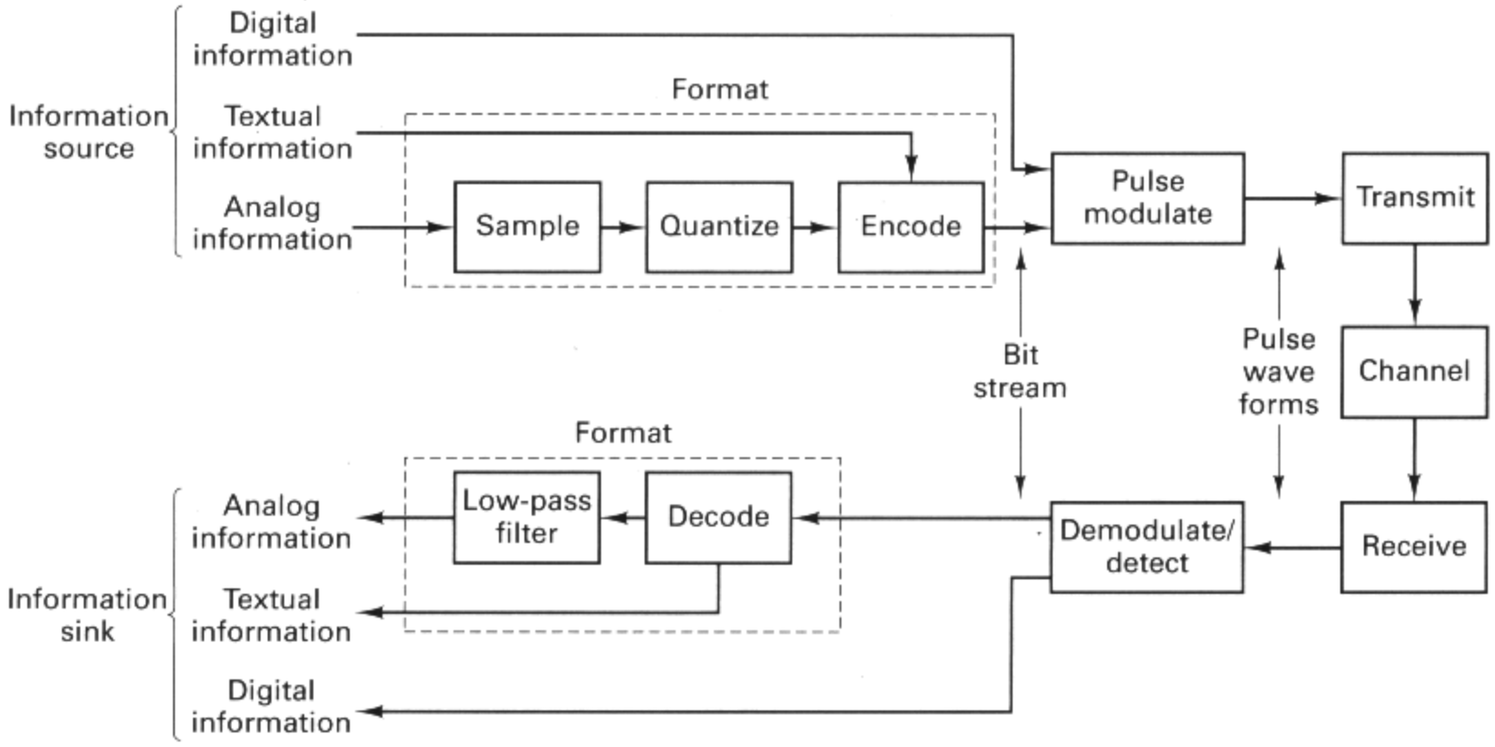


Figure 2.2 Formatting and transmission of baseband signals.

format would bypass the formatting function. Textual information is transformed into binary digits by use of a coder. Analog information is formatted using three separate processes: sampling, quantization, and coding. In all cases, the formatting step results in a sequence of binary digits.

These digits are to be transmitted through a *baseband channel,* such as a pair of wires or a coaxial cable. However, no channel can be used for the transmission of binary digits without first transforming the digits to *waveforms* that are compatible with the channel. For baseband channels, compatible waveforms are pulses.

In Figure 2.2, the conversion from a bit stream to a sequence of pulse waveforms takes place in the block labeled pulse modulate. The output of the modulator is typically a sequence of pulses with characteristics that correspond to the digits being sent. After transmission through the channel, the pulse waveforms are recovered (demodulated) and detected to produce an estimate of the transmitted digits; the final step, (reverse) formatting, recovers an estimate of the source information.

2.2 FORMATTING TEXTUAL DATA (CHARACTER CODING)

The original form of most communicated data (except for computer-to-computer transmissions) is either textual or analog. If the data consist of alphanumeric text, they will be character encoded with one of several standard formats; examples include the American Standard Code for Information Interchange (ASCII), the Extended Binary Coded Decimal Interchange Code (EBCDIC), Baudot, and Hollerith. The textual material is thereby transformed into a digital format. The ASCII format is shown in Figure 2.3; the EBCDIC format is shown in Figure 2.4.

Bits				5	0	1	0	1	0	1	0	1
				6	0	0	1	1	0	0	1	1
1	2	3	4	7	0	0	0	0	1	1	1	1
0	0	0	0		NUL	DLE	SP	0	@	P	'	p
1	0	0	0		SOH	DC1	!	1	A	Q	a	q
0	1	0	0		STX	DC2	"	2	B	R	b	r
1	1	0	0		ETX	DC3	#	3	C	S	c	s
0	0	1	0		EOT	DC4	$	4	D	T	d	t
1	0	1	0		ENQ	NAK	%	5	E	U	e	u
0	1	1	0		ACK	SYN	&	6	F	V	f	v
1	1	1	0		BEL	ETB	'	7	G	W	g	w
0	0	0	1		BS	CAN	(	8	H	X	h	x
1	0	0	1		HT	EM	)	9	I	Y	i	y
0	1	0	1		LF	SUB	*	:	J	Z	j	z
1	1	0	1		VT	ESC	+	;	K	[	k	{
0	0	1	1		FF	FS	,	<	L	\	l	|
1	0	1	1		CR	GS	–	=	M	]	m	}
0	1	1	1		SO	RS	.	>	N	^	n	~
1	1	1	1		SI	US	/	?	O	_	o	DEL

NUL	Null, or all zeros	DC1	Device control 1
SOH	Start of heading	DC2	Device control 2
STX	Start of text	DC3	Device control 3
ETX	End of text	DC4	Device control 4
EOT	End of transmission	NAK	Negative acknowledge
ENQ	Enquiry	SYN	Synchronous idle
ACK	Acknowledge	ETB	End of transmission
BEL	Bell, or alarm	CAN	Cancel
BS	Backspace	EM	End of medium
HT	Horizontal tabulation	SUB	Substitute
LF	Line feed	ESC	Escape
VT	Vertical tabulation	FS	File separator
FF	Form feed	GS	Group separator
CR	Carriage return	RS	Record separator
SO	Shift out	US	Unit separator
SI	Shift in	SP	Space
DLE	Data link escape	DEL	Delete

Figure 2.3 Seven-bit American standard code for information interchange (ASCII).

Bits				5	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
				6	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
				7	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
				8	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
1	2	3	4																	
0	0	0	0		NUL	SOH	STX	ETX	PF	HT	LC	DEL			SMM	VT	FF	CR	SO	SI
0	0	0	1		DLE	DC1	DC2	DC3	RES	NL	BS	IL	CAN	EM	CC		IFS	IGS	IRS	IUS
0	0	1	0		DS	SOS	FS		BYP	LF	EOB	PRE			SM			ENQ	ACK	BEL
0	0	1	1				SYN		PN	RS	US	EOT					DC4	NAK		SUB
0	1	0	0		SP										¢		<	(	+	!
0	1	0	1		&										!	$	*	)	;	¬
0	1	1	0		–	/										,	%	_	>	?
0	1	1	1												:	#	@	'	=	"
1	0	0	0			a	b	c	d	e	f	g	h	i						
1	0	0	1			j	k	l	m	n	o	p	q	r						
1	0	1	0				s	t	u	v	w	x	y	z						
1	0	1	1																	
1	1	0	0			A	B	C	D	E	F	G	H	I						
1	1	0	1			J	K	L	M	N	O	P	Q	R						
1	1	1	0				S	T	U	V	W	X	Y	Z						
1	1	1	1		0	1	2	3	4	5	6	7	8	9						

PF	Punch off
HT	Horizontal tab
LC	Lower case
DEL	Delete
SP	Space
UC	Upper case
RES	Restore
NL	New line
BS	Backspace
IL	Idle
PN	Punch on
EOT	End of transmission
BYP	Bypass
LF	Line feed
EOB	End of block
PRE	Prefix (ESC)
RS	Reader stop
SM	Start message
DS	Digit select
SOS	Start of significance
IFS	Interchange file separator
IGS	Interchange group separator
IRS	Interchange record separator
IUS	Interchange unit separator
Others	Same as ASCII

Figure 2.4 EBCDIC character code set.

The bit numbers signify the order of serial transmission, where bit number 1 is the first signaling element. Character coding, then, is the step that transforms text into binary digits (bits). Sometimes existing character codes are modified to meet specialized needs. For example, the 7-bit ASCII code (Figure 2.3) can be modified to include an added bit for error detection purposes. (See Chapter 6.) On the other hand, sometimes the code is truncated to a 6-bit ASCII version, which provides capability for only 64 characters instead of the 128 characters allowed by 7-bit ASCII.

2.3 MESSAGES, CHARACTERS, AND SYMBOLS

Textual messages comprise a sequence of alphanumeric characters. When digitally transmitted, the characters are first encoded into a sequence of bits, called a *bit stream* or *baseband signal.* Groups of k bits can then be combined to form new digits, or *symbols,* from a finite symbol set or alphabet of $M = 2^k$ such symbols. A system using a symbol set size of M is referred to as an *M-ary system.* The value of k or M represents an important initial choice in the design of any digital communication system. For $k = 1$, the system is termed *binary,* the size of the symbol set is $M = 2$, and the modulator uses one of the two different waveforms to represent the binary "one" and the other to represent the binary "zero." For this special case, the symbol and the bit are the same. For $k = 2$, the system is termed *quaternary* or 4-*ary* ($M = 4$). At each symbol time, the modulator uses one of the four different waveforms that represents the symbol. The partitioning of the sequence of message bits is determined by the specification of the symbol set size, M. The following example should help clarify the relationship between the following terms: "message," "character," "symbol," "bit," and "digital waveform."

2.3.1 Example of Messages, Characters, and Symbols

Figure 2.5 shows examples of bit stream partitioning, based on the system specification for the values of k and M. The textual message in the figure is the word "THINK." Using 6-bit ASCII character coding (bit numbers 1 to 6 from Figure 2.3) yields a bit stream comprising 30 bits. In Figure 2.5a, the symbol set size, M, has been chosen to be 8 (each symbol represents an 8-ary digit). The bits are therefore partitioned into groups of three ($k = \log_2 8$); the resulting 10 numbers represent the 10 octal symbols to be transmitted. The transmitter must have a repertoire of eight waveforms $s_i(t)$, where $i = 1, \ldots, 8$, to represent the possible symbols, any one of which may be transmitted during a symbol time. The final row of Figure 2.5a lists the 10 waveforms that an 8-ary modulating system transmits to represent the textual message "THINK."

In Figure 2.5b, the symbol set size, M, has been chosen to be 32 (each symbol represents a 32-ary digit). The bits are therefore taken five at a time, and the resulting group of six numbers represent the six 32-ary symbols to be transmitted. Notice that there is no need for the symbol boundaries and the character boundaries to coincide. The first symbol represents $\frac{5}{6}$ of the first character, "T." The second symbol

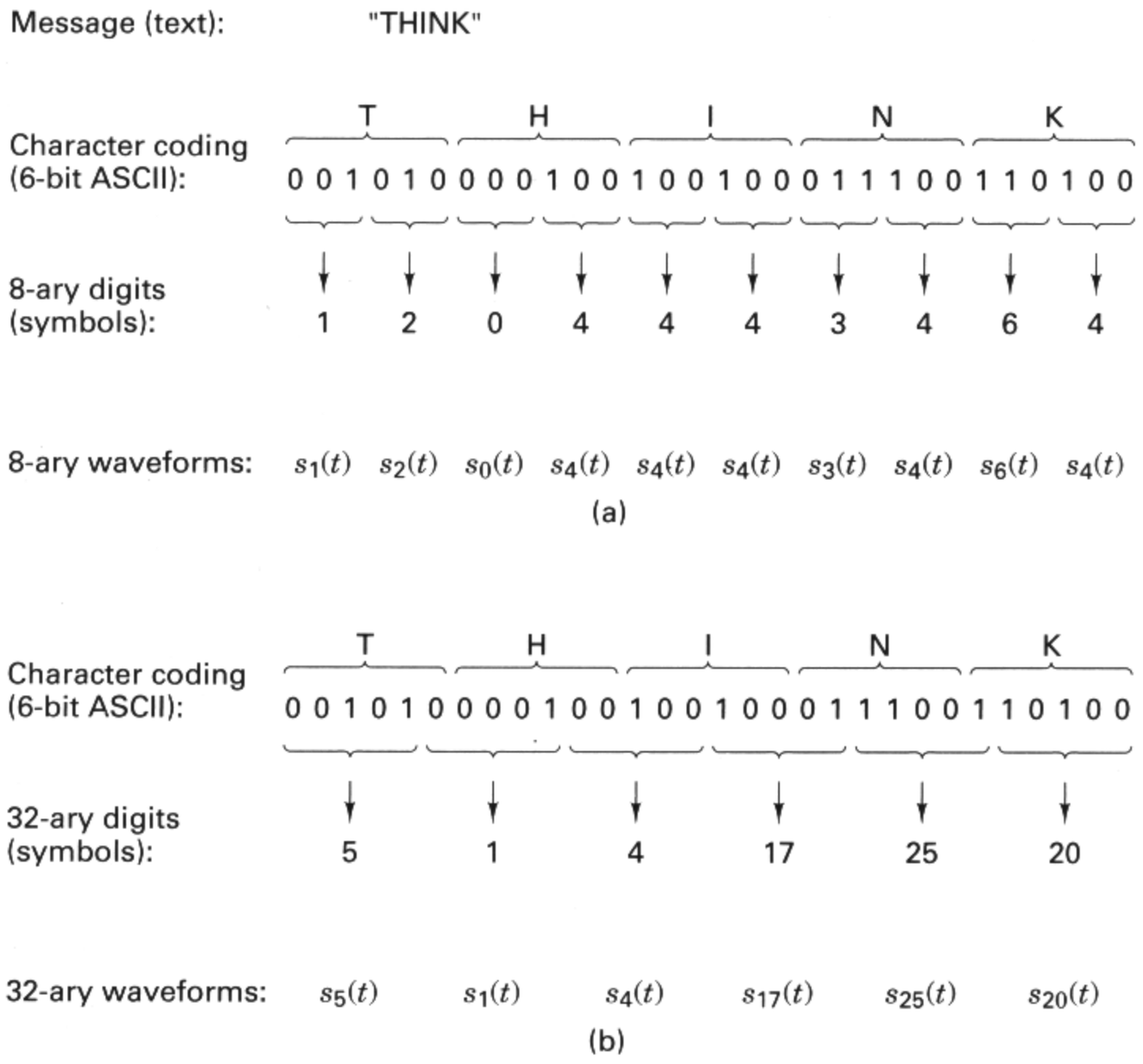


Figure 2.5 Messages, characters, and symbols. (a) 8-ary example. (b) 32-ary example.

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

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

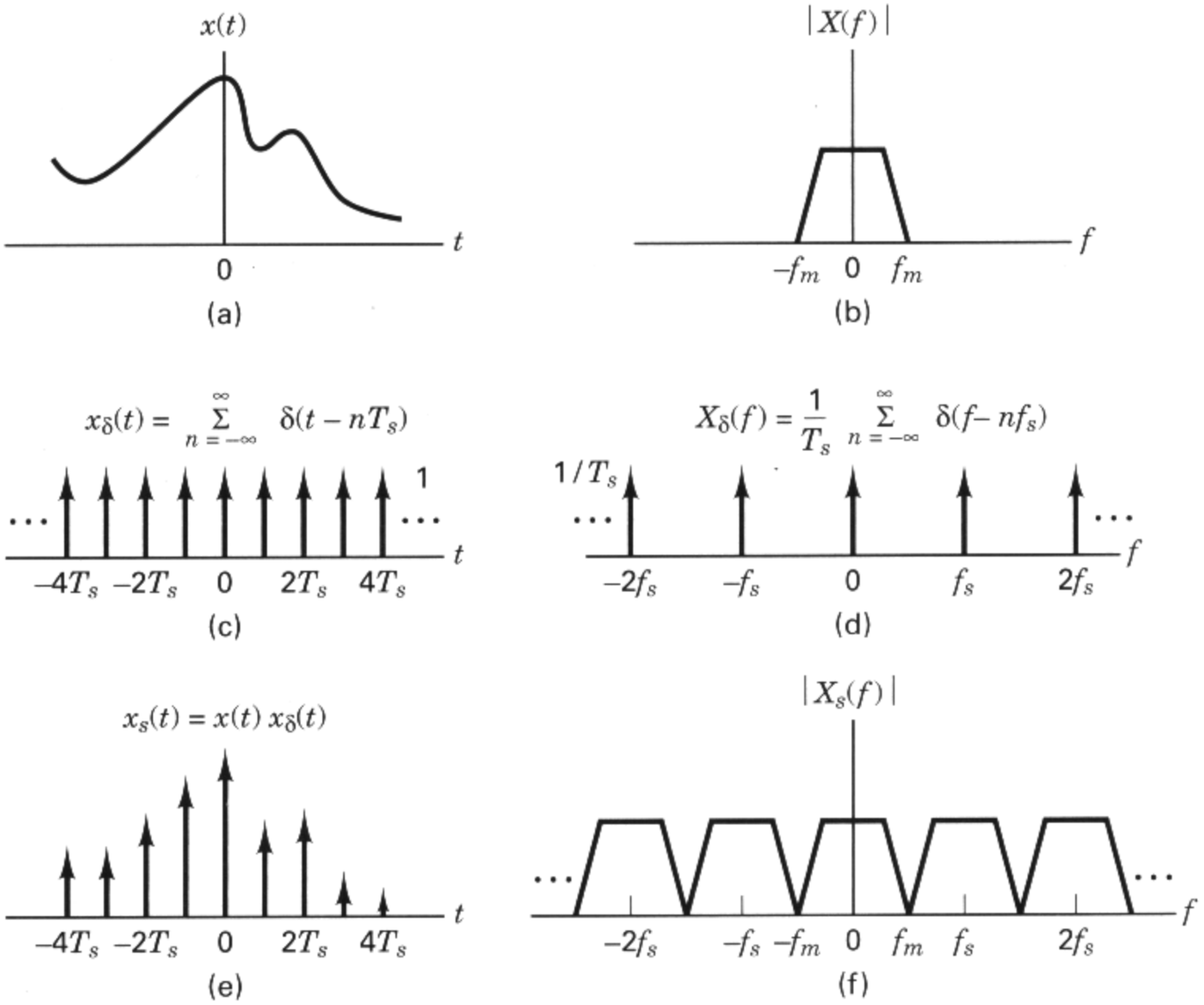


Figure 2.6 Sampling theorem using the frequency convolution property of the Fourier transform.

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

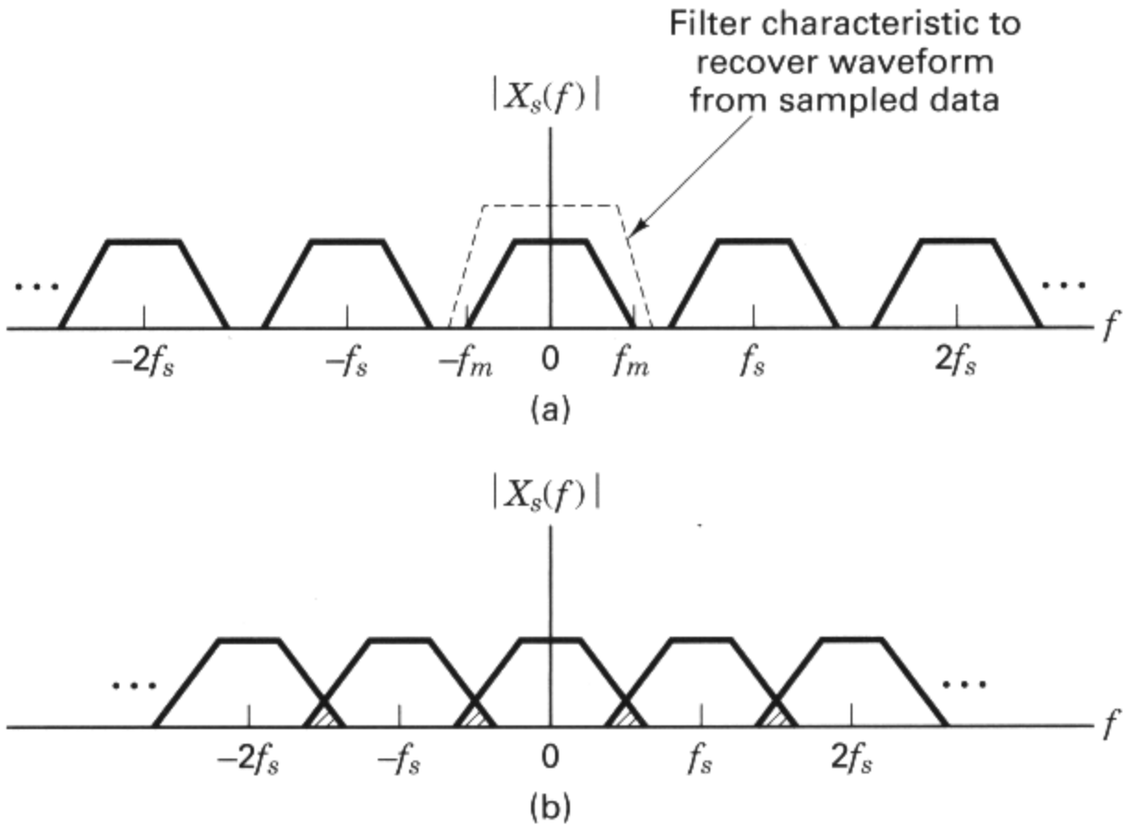


Figure 2.7 Spectra for various sampling rates. (a) Sampled spectrum ($f_s > 2f_m$). (b) Sampled spectrum ($f_s < 2f_m$).

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

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

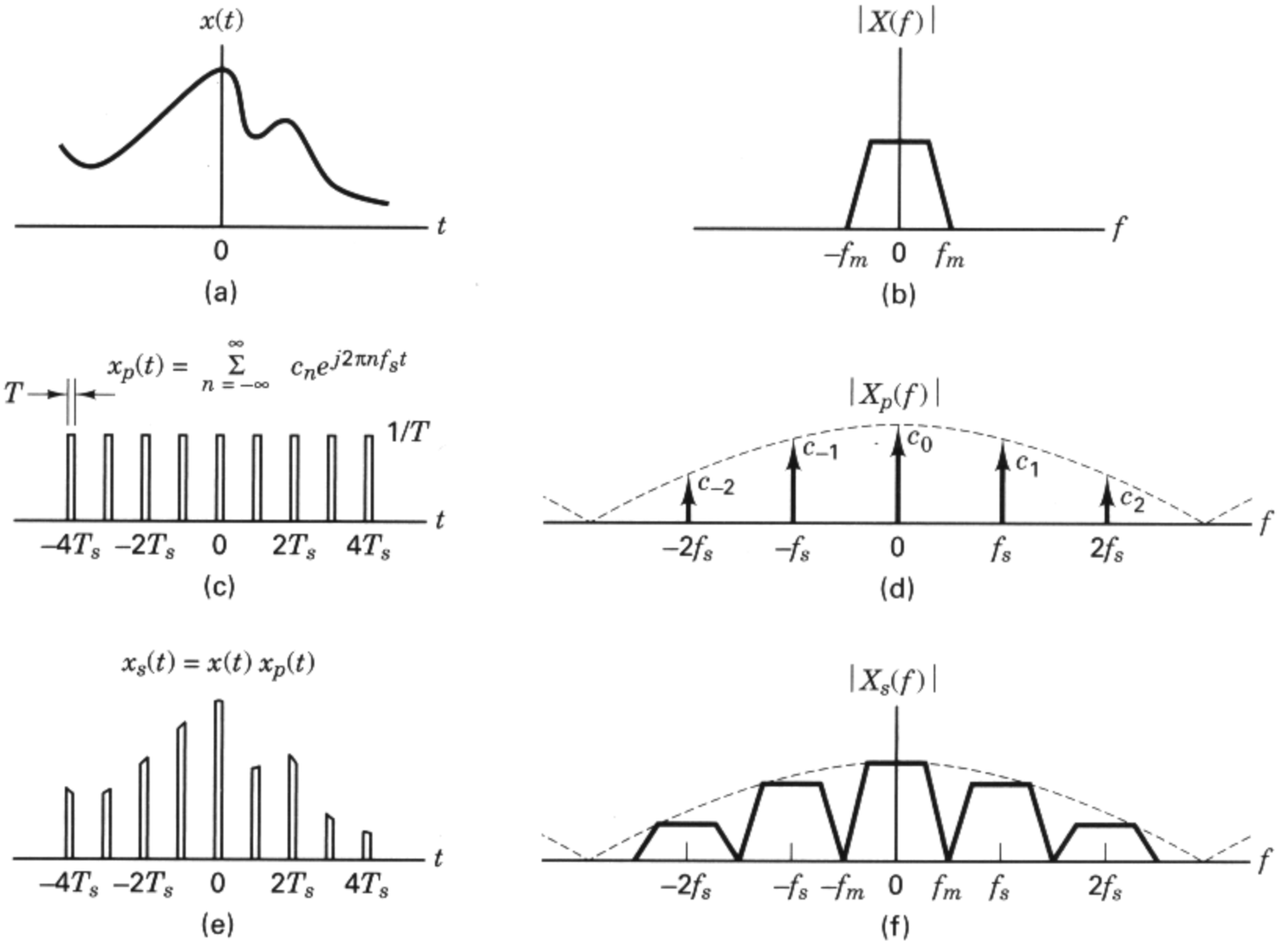


Figure 2.8 Sampling theorem using the frequency shifting property of the Fourier transform.

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-

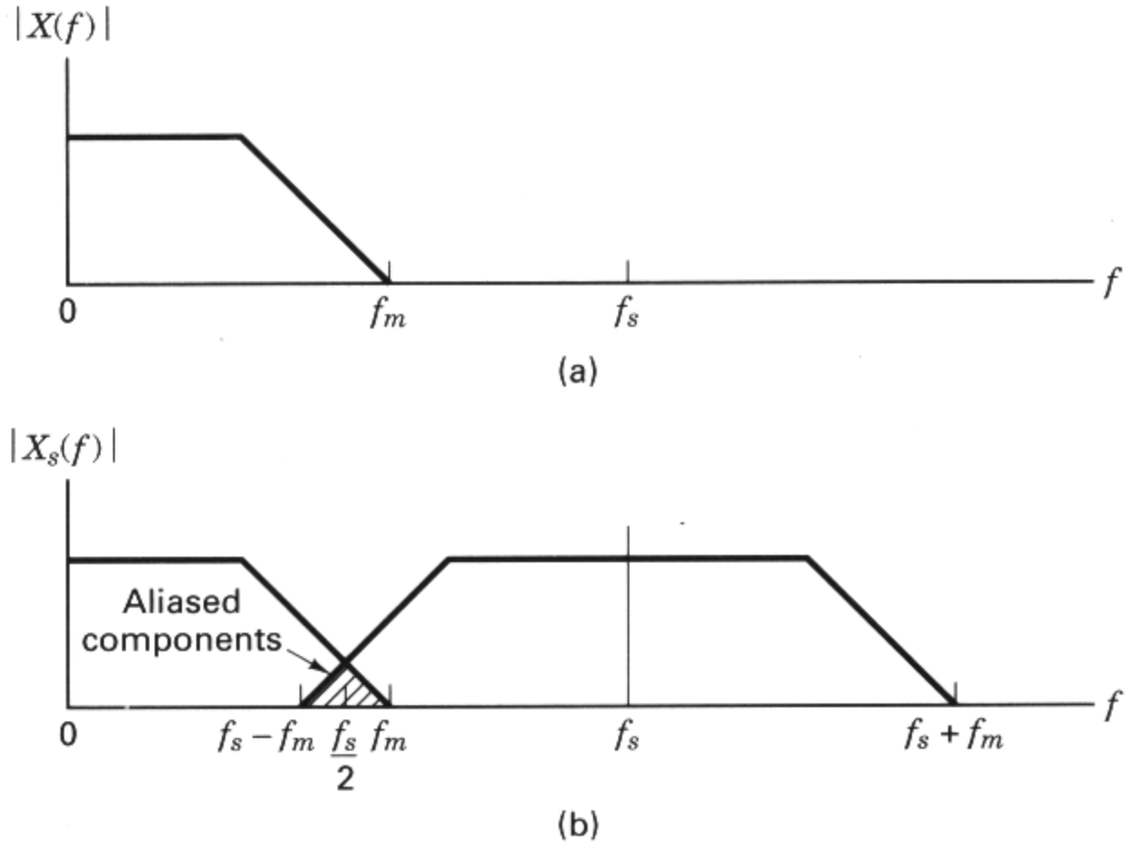


Figure 2.9 Aliasing in the frequency domain. (a) Continuous signal spectrum. (b) Sampled signal spectrum.

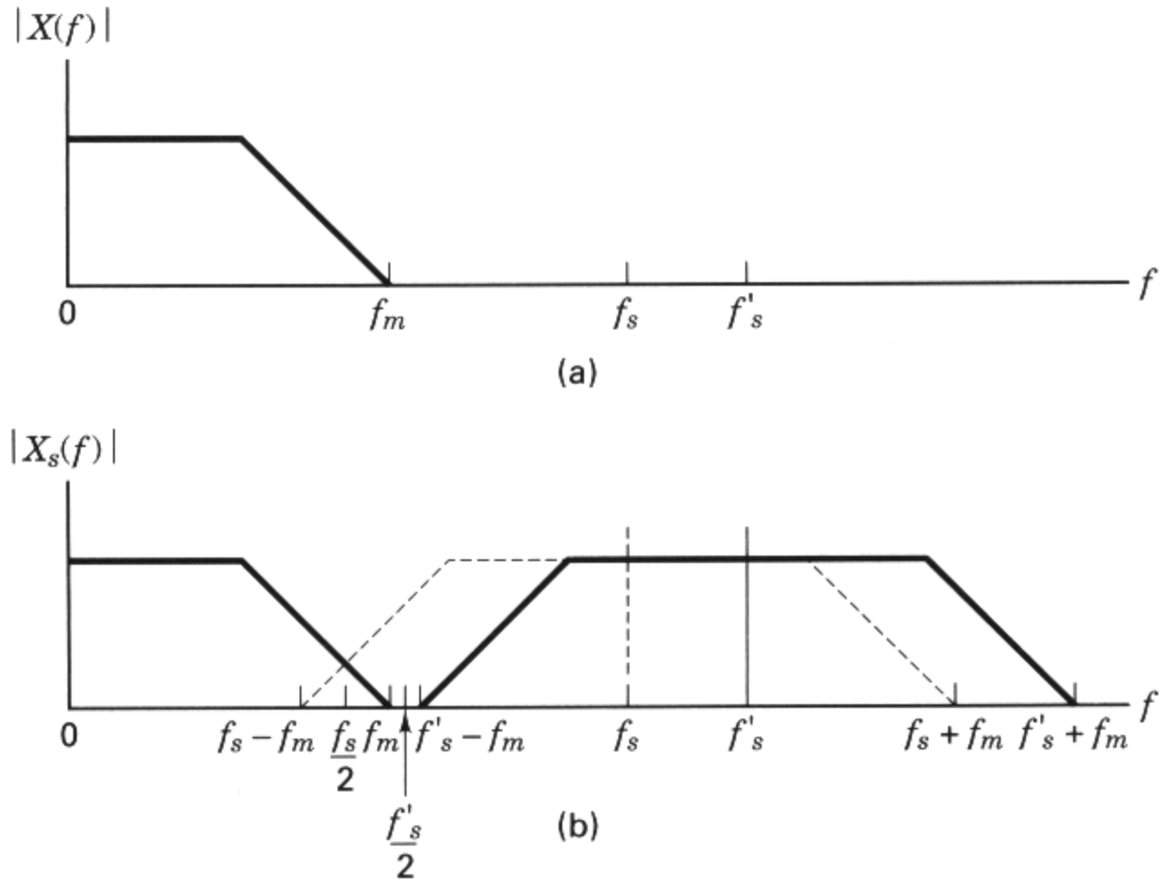


Figure 2.10 Higher sampling rate eliminates aliasing. (a) Continuous signal spectrum. (b) Sampled signal spectrum.

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-

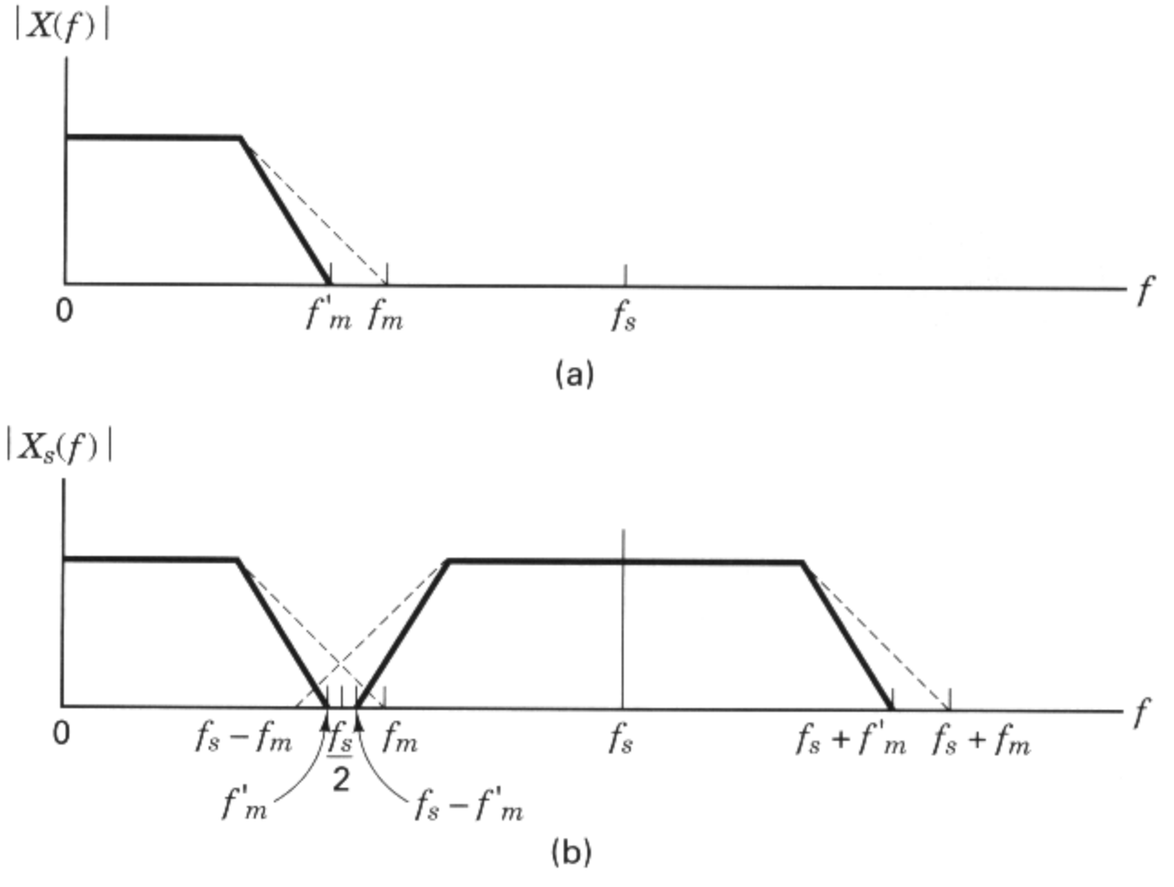


Figure 2.11 Sharper-cutoff filters eliminate aliasing. (a) Continuous signal spectrum. (b) Sampled signal spectrum.

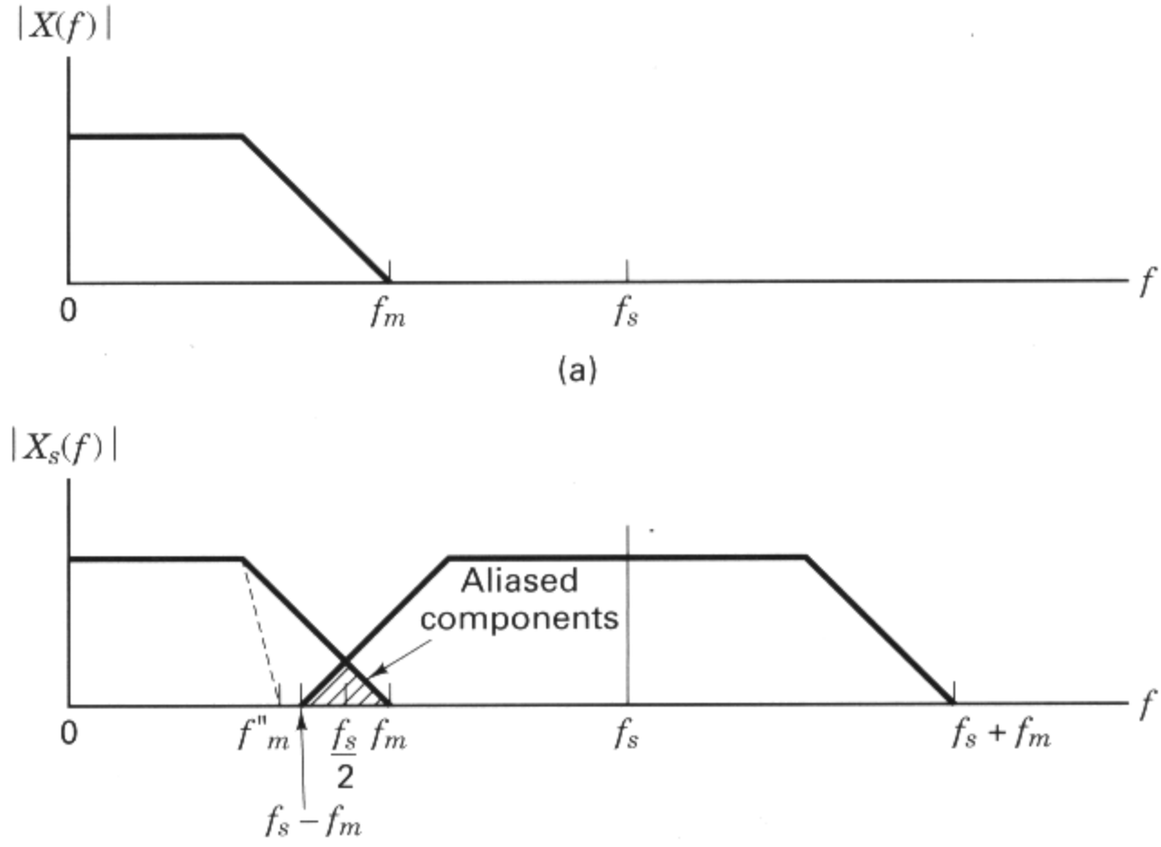


Figure 2.12 Postfilter eliminates aliased portion of spectrum. (a) Continuous signal spectrum. (b) Sampled signal spectrum.

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

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

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-

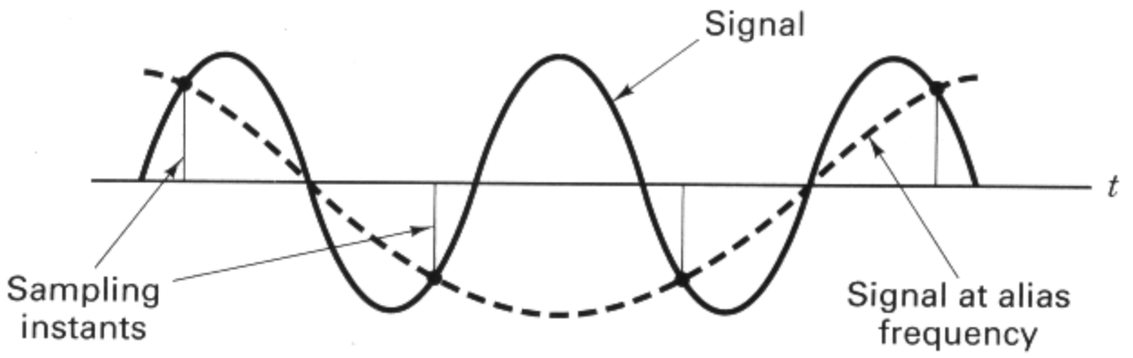


Figure 2.13 Alias frequency generated by sub-Nyquist sampling rate.

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).

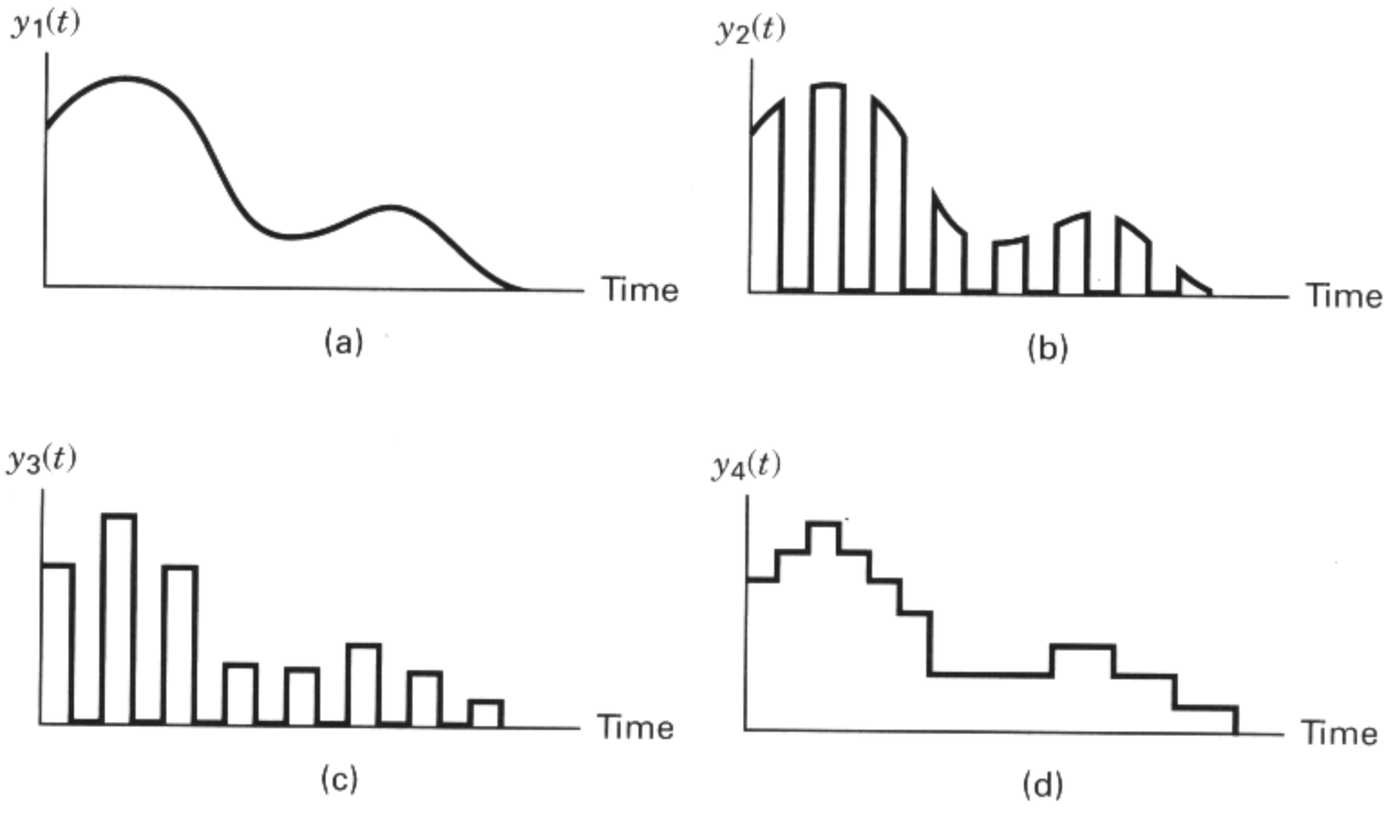


Figure 2.14 Amplitude and time coordinates of source data. (a) Original analog waveform. (b) Natural-sampled data. (c) Quantized samples. (d) Sample and hold.

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

called the *operating range* of the converter. If the input exceeds this range, the difference between the input and the output becomes large, and we say that the converter is operating in *saturation.* Saturation errors, being large, are more objectionable than quantizing noise. Generally, saturation is avoided by the use of automatic gain control (AGC), which effectively extends the operating range of the converter. (Chapter 13 covers quantizer saturation in greater detail.)

2.5.1.3 Timing Jitter

Our analysis of the sampling theorem predicted precise reconstruction of the signal based on uniformly spaced samples of the signal. If there is a slight jitter in the position of the sample, the sampling is no longer uniform. Although exact reconstruction is still possible if the sample positions are accurately known, the jitter is usually a random process and thus the sample positions are not accurately known. The effect of the jitter is equivalent to frequency modulation (FM) of the baseband signal. If the jitter is random, a low-level wideband spectral contribution is induced whose properties are very close to those of the quantizing noise. If the jitter exhibits periodic components, as might be found in data extracted from a tape recorder, the periodic FM will induce low-level spectral lines in the data. Timing jitter can be controlled with very good power supply isolation and stable clock references.

2.5.2 Channel Effects

2.5.2.1 Channel Noise

Thermal noise, interference from other users, and interference from circuit switching transients can cause errors in detecting the pulses carrying the digitized samples. Channel-induced errors can degrade the reconstructed signal quality quite quickly. This rapid degradation of output signal quality with channel-induced errors is called a *threshold effect.* If the channel noise is small, there will be no problem detecting the presence of the waveforms. Thus, small noise does not corrupt the reconstruct signals. In this case, the only noise present in the reconstruction is the quantization noise. On the other hand, if the channel noise is large enough to affect our ability to detect the waveforms, the resulting detection error causes reconstruction errors. A large difference in behavior can occur for very small changes in channel noise level.

2.5.2.2 Intersymbol Interference

The channel is always bandlimited. A bandlimited channel disperses or spreads a pulse waveform passing through it (see Section 1.6.4). When the channel bandwidth is much greater than the pulse bandwidth, the spreading of the pulse will be slight. When the channel bandwidth is close to the signal bandwidth, the spreading will exceed a symbol duration and cause signal pulses to overlap. This overlapping is called *intersymbol interference* (ISI). Like any other source of interference, ISI causes system degradation (higher error rates); it is a particularly

insidious form of interference because raising the signal power to overcome the interference will not always improve the error performance. (Details of how ISI is handled are presented in the next chapter, in Sections 3.3 and 3.4.)

2.5.3 Signal-to-Noise Ratio for Quantized Pulses

Figure 2.15 illustrates an L-level linear quantizer for an analog signal with a peak-to-peak voltage range of $V_{pp} = V_p - (-V_p) = 2V_p$ volts. The quantized pulses assume positive and negative values, as shown in the figure. The step size between quantization levels, called the *quantile interval,* is denoted q volts. When the quantization levels are uniformly distributed over the full range, the quantizer is called a *uniform or linear quantizer.* Each sample value of the analog waveform is approximated with a quantized pulse; the approximation will result in an error no larger than $q/2$ in the positive direction or $-q/2$ in the negative direction. The degradation of the signal due to quantization is therefore limited to half a quantile interval, $\pm q/2$ volts.

A useful figure of merit for the uniform quantizer is the quantizer variance (mean-square error assuming zero mean). If we assume that the quantization error, e, is uniformly distributed over a single quantile interval q-wide (i.e., the analog input takes on all values with equal probability), the quantizer error variance is found to be

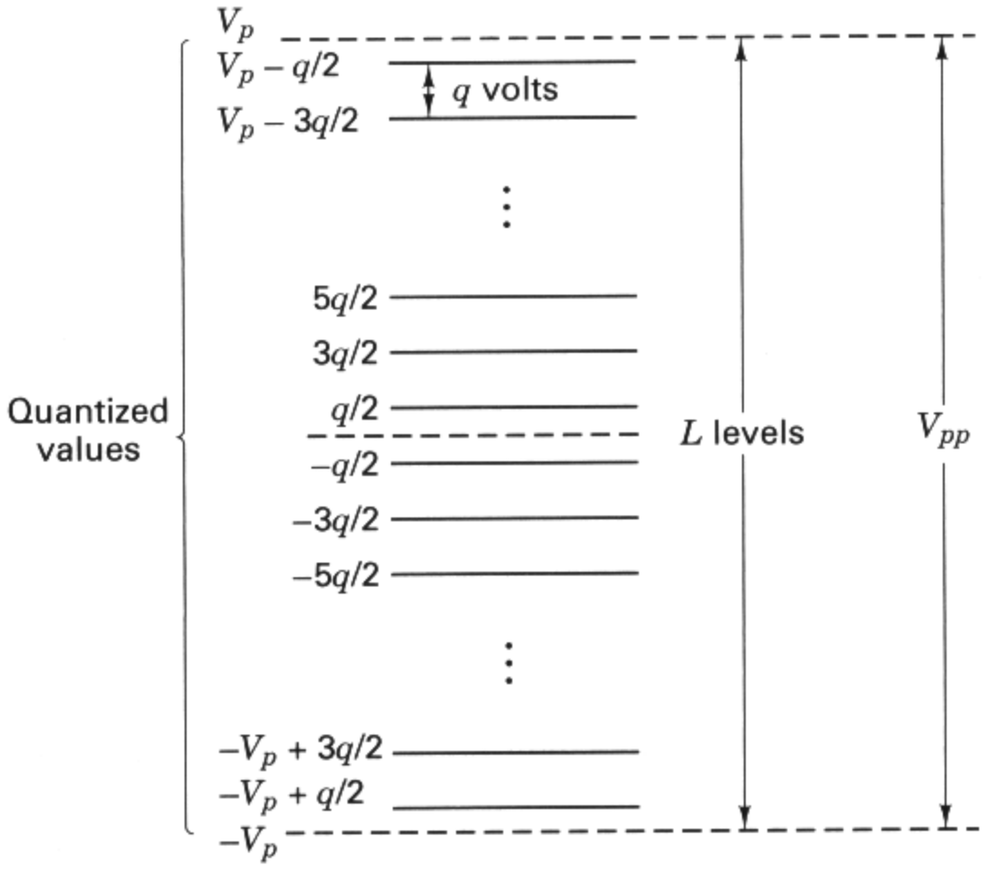


Figure 2.15 Quantization levels.

$$\sigma^2 = \int_{-q/2}^{+q/2} e^2 p(e)\, de \tag{2.18a}$$

$$= \int_{-q/2}^{+q/2} e^2 \frac{1}{q}\, de = \frac{q^2}{12} \tag{2.18b}$$

where $p(e) = 1/q$ is the (uniform) probability density function of the quantization error. The variance, σ^2, corresponds to the *average quantization noise power.* The peak power of the analog signal (normalized to 1 Ω) can be expressed as

$$V_p^2 = \left(\frac{V_{pp}}{2}\right)^2 = \left(\frac{Lq}{2}\right)^2 = \frac{L^2q^2}{4} \tag{2.19}$$

where L is the number of quantization levels. Equations (2.18) and (2.19) combined yield the ratio of *peak* signal power to *average* quantization noise power $(S/N)_q$, assuming that there are no errors due to ISI or channel noise:

$$\left(\frac{S}{N}\right)_q = \frac{L^2q^2/4}{q^2/12} = 3L^2 \tag{2.20}$$

It is intuitively satisfying to see that $(S/N)_q$ improves as a function of the number of quantization levels squared. In the limit (as $L \to \infty$), the signal approaches the PAM format (with no quantization), and the signal-to-quantization noise ratio is infinite; in other words, with an infinite number of quantization levels, there is zero quantization noise.

2.6 PULSE CODE MODULATION

Pulse code modulation (PCM) is the name given to the class of baseband signals obtained from the quantized PAM signals by encoding each quantized sample into a *digital word* [3]. The source information is sampled and quantized to one of L levels; then each quantized sample is digitally encoded into an ℓ-bit ($\ell = \log_2 L$) codeword. For baseband transmission, the codeword bits will then be transformed to pulse waveforms. The essential features of binary PCM are shown in Figure 2.16. Assume that an analog signal $x(t)$ is limited in its excursions to the range –4 to +4 V. The step size between quantization levels has been set at 1 V. Thus, eight quantization levels are employed; these are located at –3.5, –2.5, . . . , +3.5 V. We assign the code number 0 to the level at –3.5 V, the code number 1 to the level at –2.5 V, and so on, until the level at 3.5 V, which is assigned the code number 7. Each code number has its representation in binary arithmetic, ranging from 000 for code number 0 to 111 for code number 7. Why have the voltage levels been chosen in this manner, compared with using a sequence of consecutive integers, 1, 2, 3, . . . ? The choice of voltage levels is guided by two constraints. First, the quantile intervals between the levels should be equal; and second, it is convenient for the levels to be symmetrical about zero.

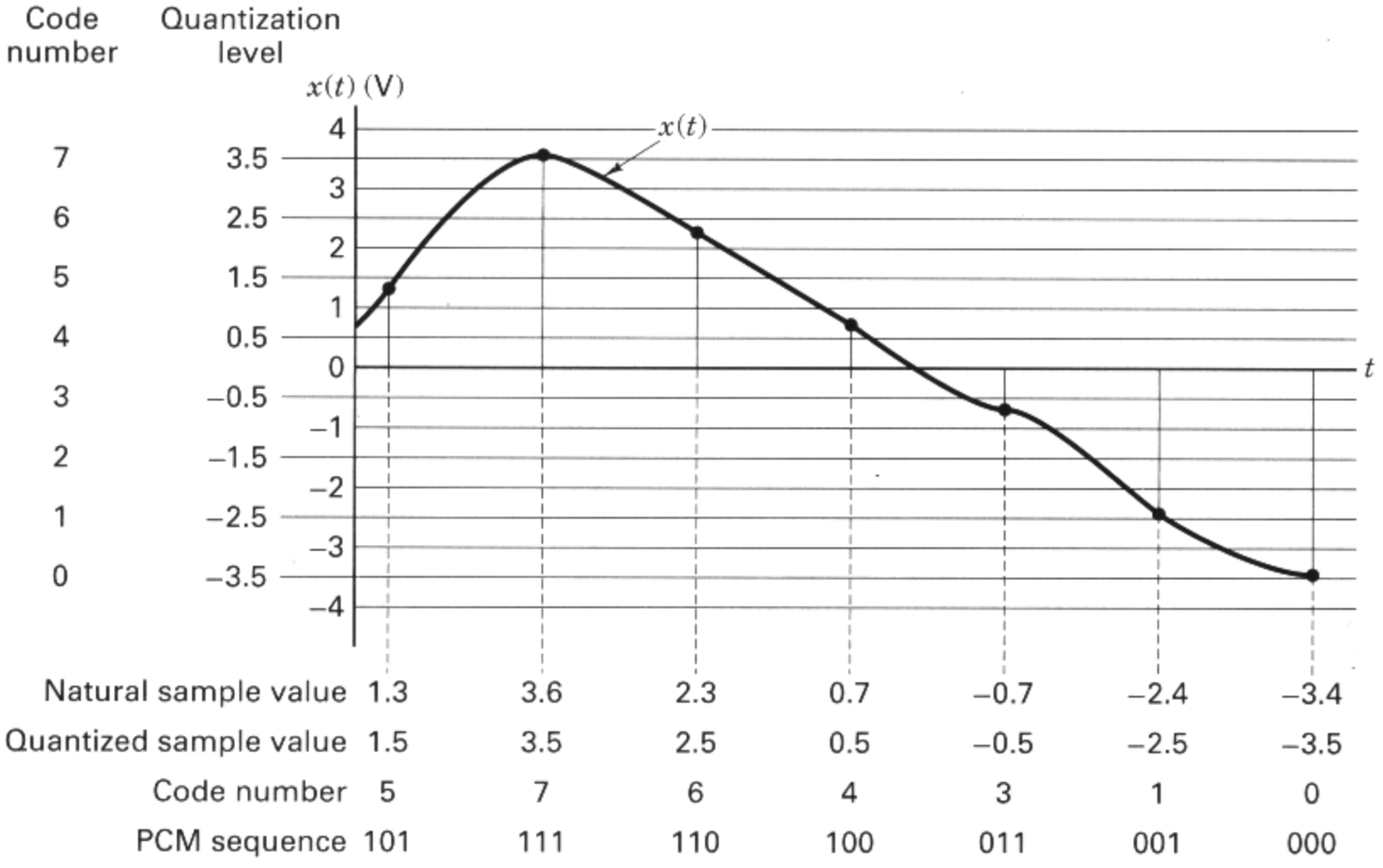


Figure 2.16 Natural samples, quantized samples, and pulse code modulation. (Reprinted with permission from Taub and Schilling, *Principles of Communications Systems,* McGraw-Hill Book Company, New York, 1971, Fig. 6.5-1, p. 205.)

The ordinate in Figure 2.16 is labeled with quantization levels and their code numbers. Each sample of the analog signal is assigned to the quantization level closest to the value of the sample. Beneath the analog waveform $x(t)$ are seen four representations of $x(t)$, as follows: the natural sample values, the quantized sample values, the code numbers, and the PCM sequence.

Note, that in the example of Figure 2.16, each sample is assigned to one of eight levels or a three-bit PCM sequence. Suppose that the analog signal is a musical passage, which is sampled at the Nyquist rate. And, suppose that when we listen to the music in digital form, it sounds terrible. What could we do to improve the fidelity? Recall that the process of quantization replaces the true signal with an approximation (i.e., adds quantization noise). Thus, increasing the number of levels will reduce the quantization noise. If we double the number of levels to 16, what are the consequences? In that case, each analog sample will be represented as a four-bit PCM sequence. Will that cost anything? In a real-time communication system, the messages must not be delayed. Hence, the transmission time for each sample must be the same, regardless of how many bits represent the sample. Hence, when there are more bits per sample, the bits must move faster; in other words, they must be replaced by "skinnier" bits. The data rate is thus increased, and the cost is a greater transmission bandwidth. This explains how one can generally obtain better fidelity at the cost of more transmission bandwidth. Be aware, however,

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

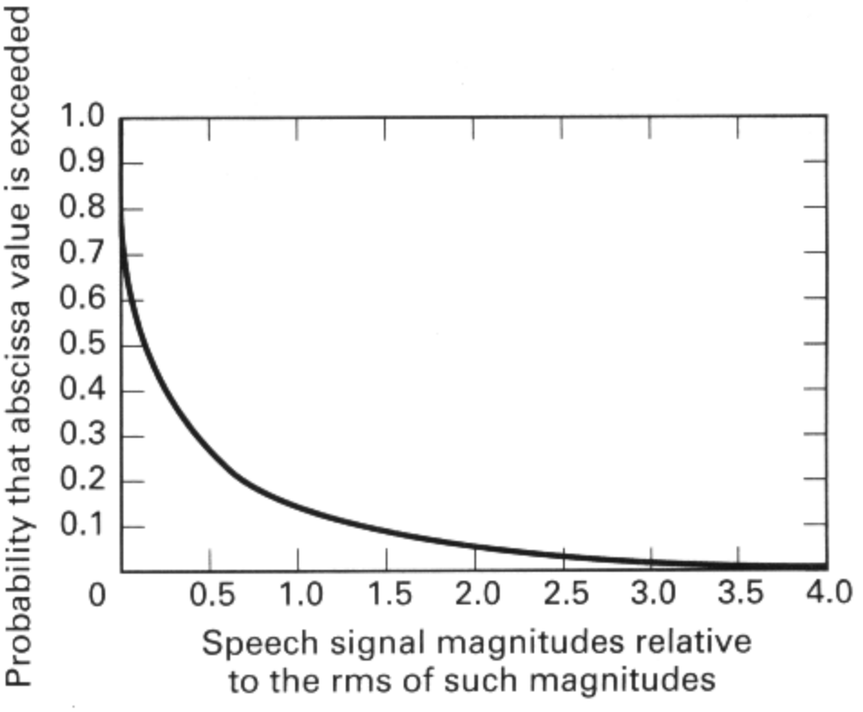


Figure 2.17 Statistical distribution of single-talker speech signal magnitudes.

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

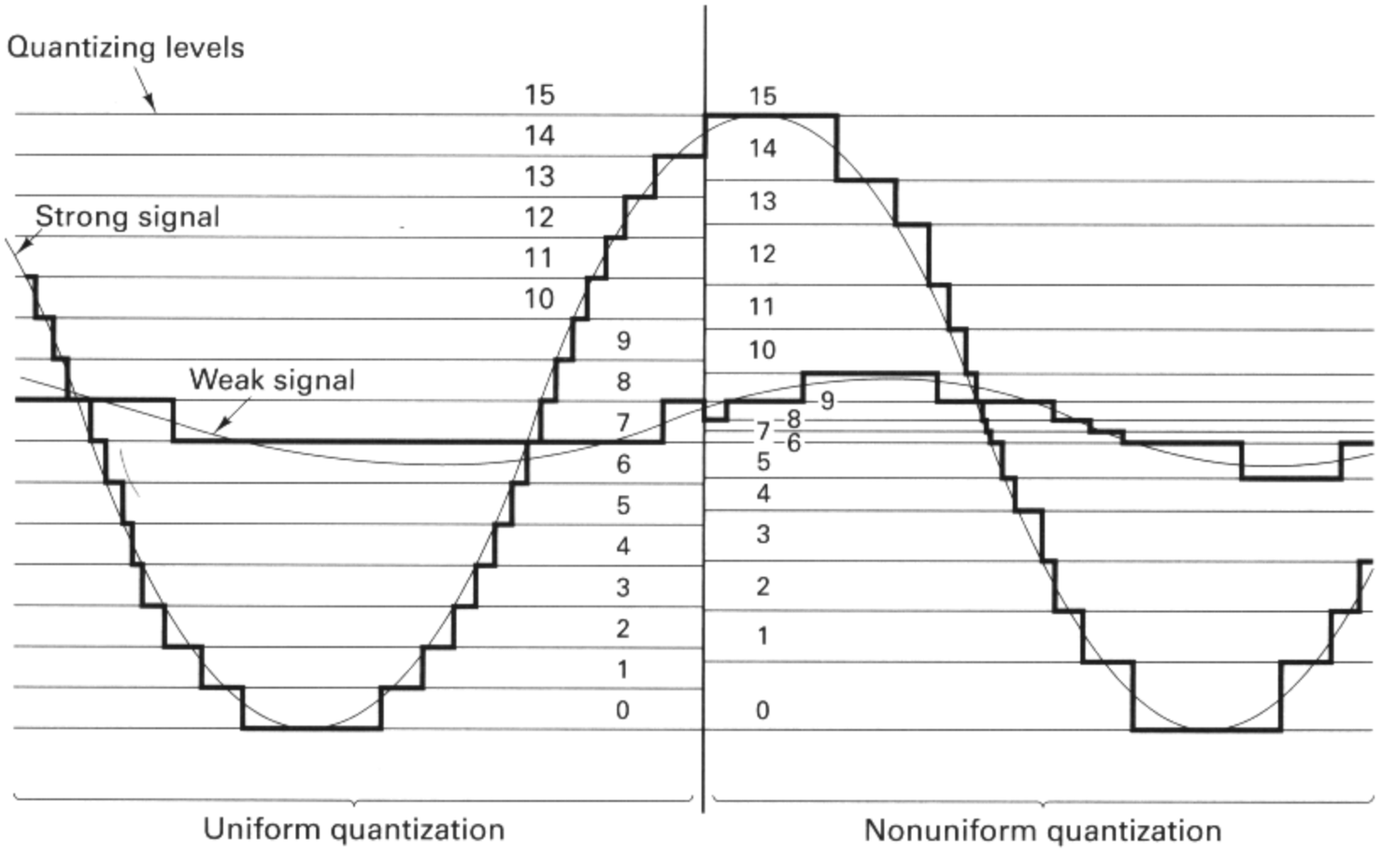


Figure 2.18 Uniform and nonuniform quantization of signals.

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

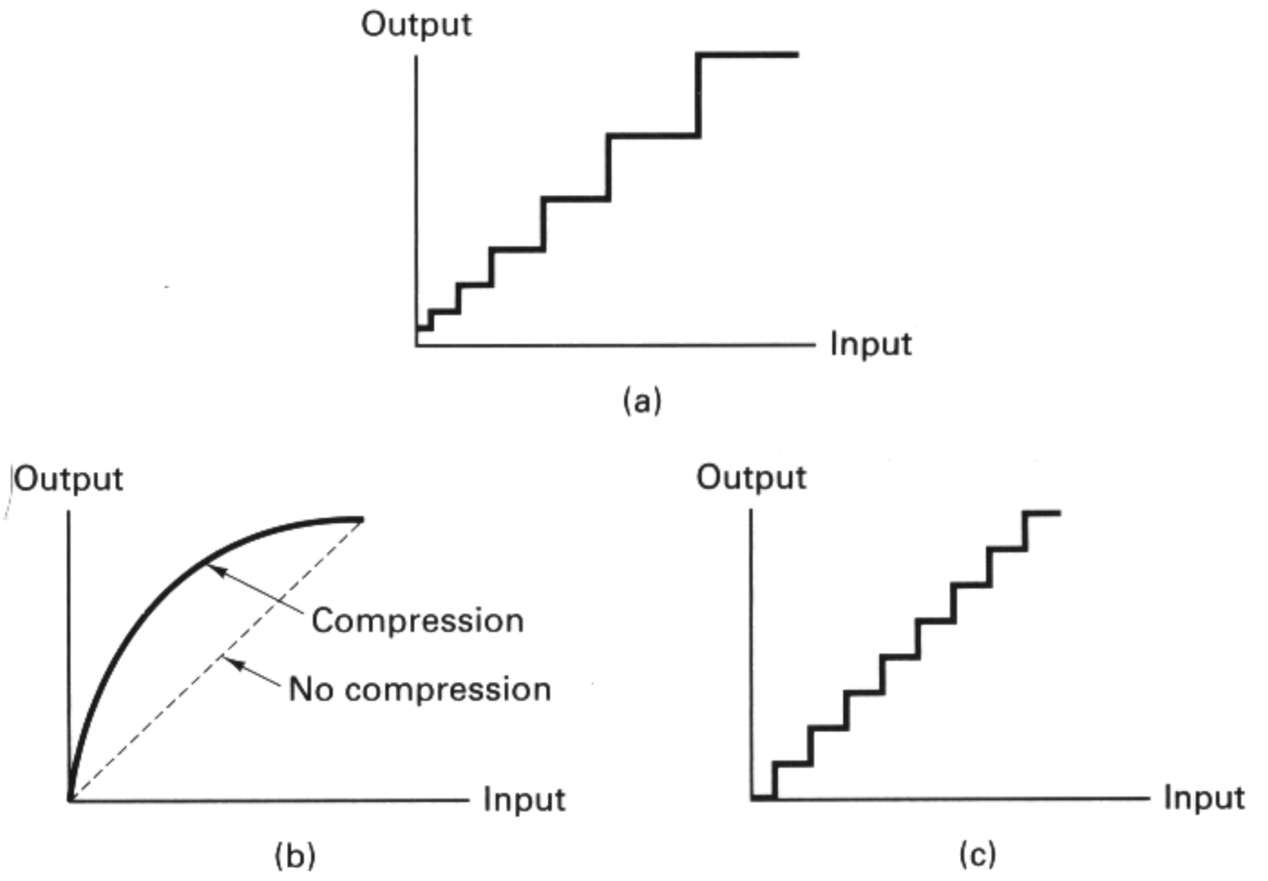


Figure 2.19 (a) Nonuniform quantizer characteristic. (b) Compression characteristic. (c) Uniform quantizer characteristic.

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

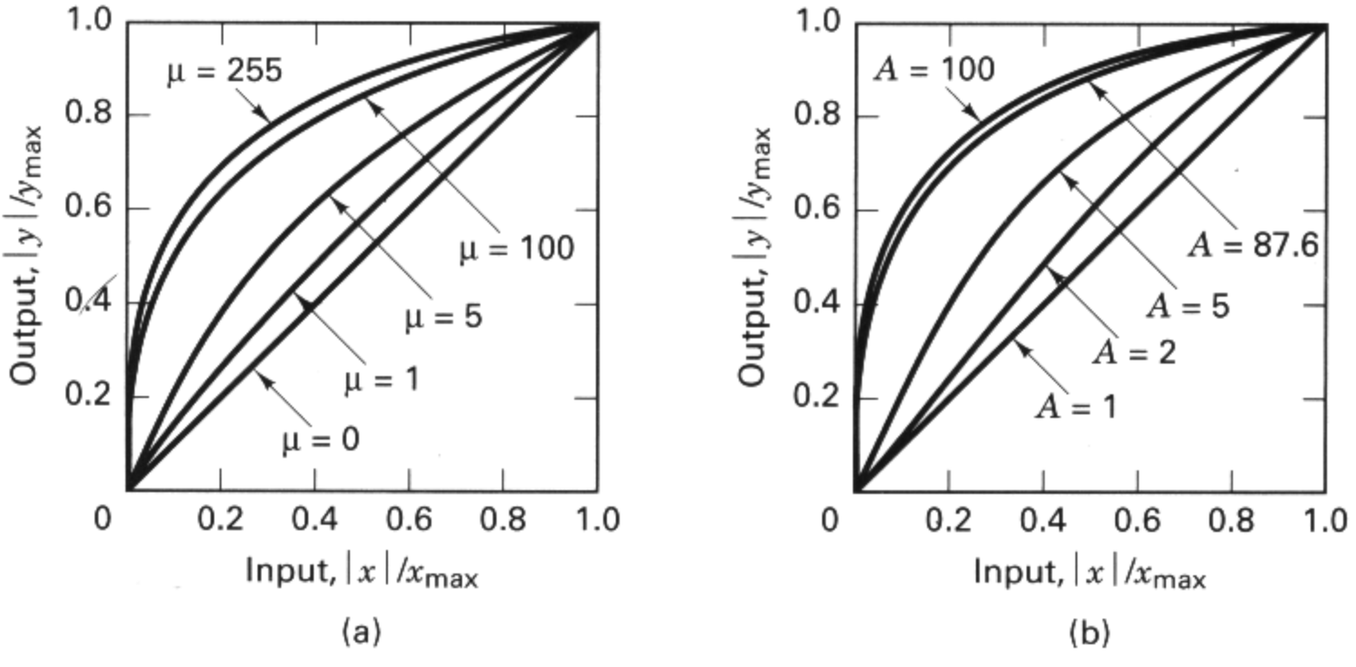


Figure 2.20 Compression characteristics. (a) μ-law characteristic. (b) A-law characteristic.

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

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

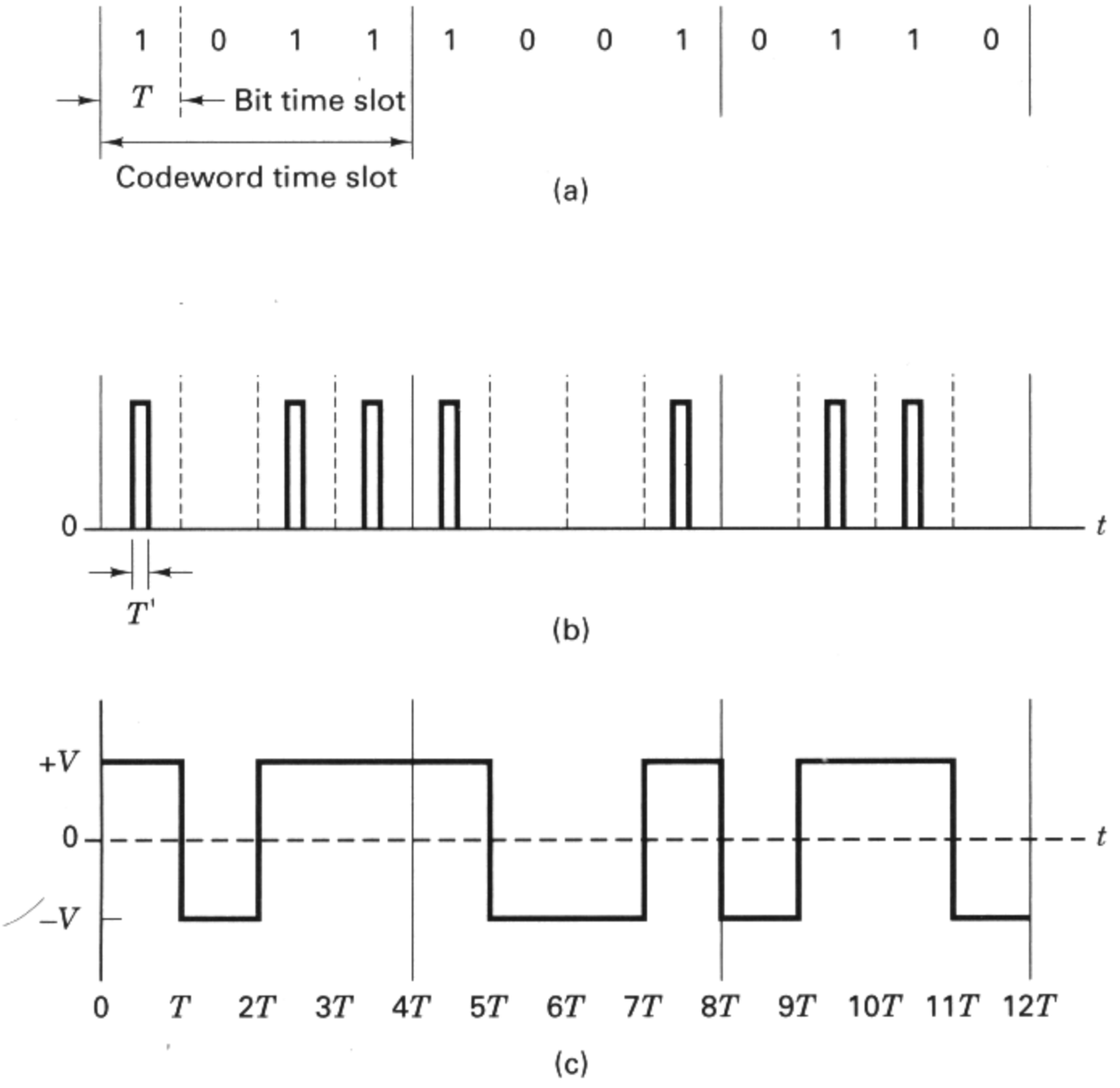


Figure 2.21 Example of waveform representation of binary digits. (a) PCM sequence. (b) Pulse representation of PCM. (c) Pulse waveform (transition between two levels).

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.

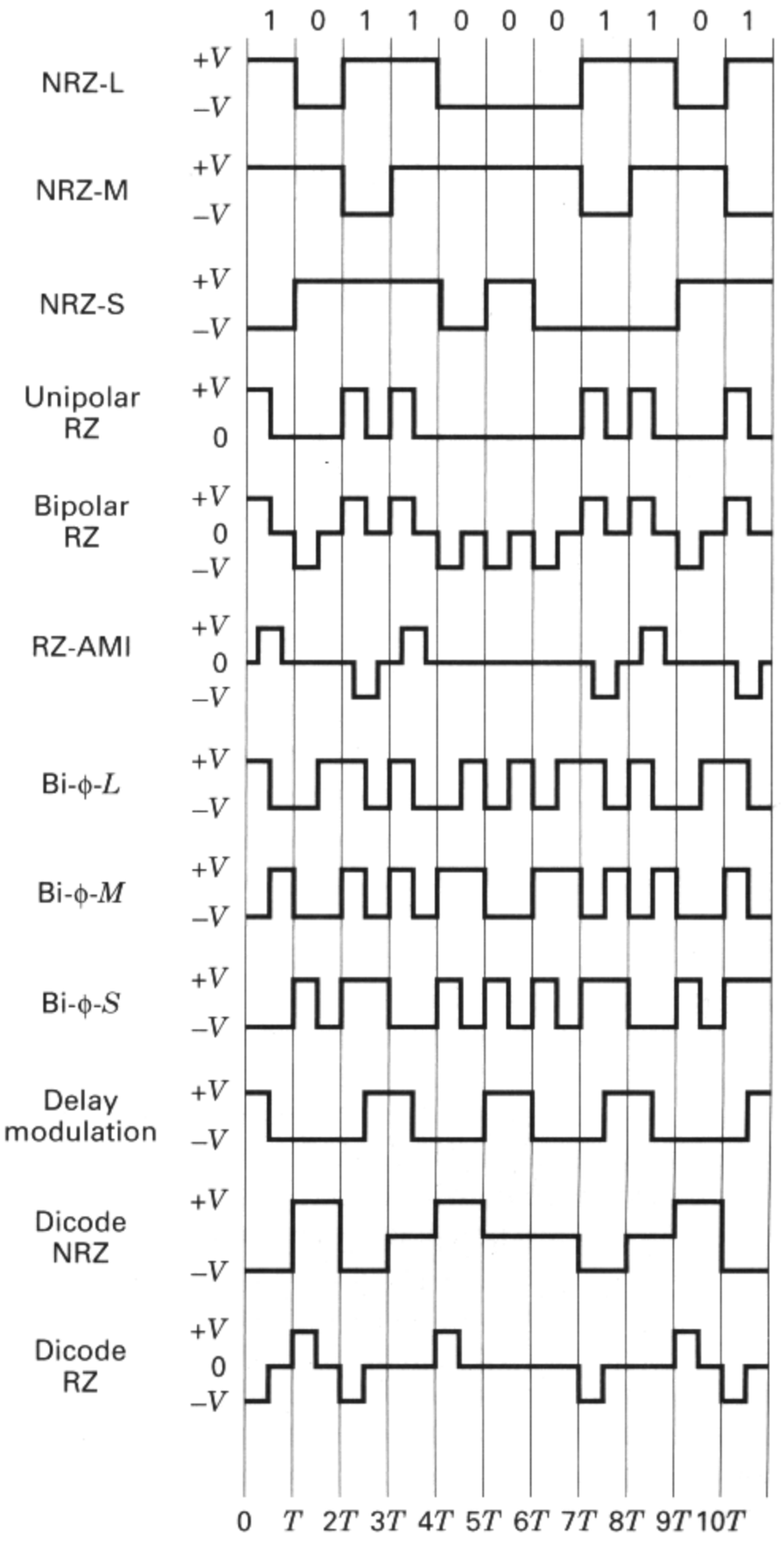


Figure 2.22 Various PCM waveforms.

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

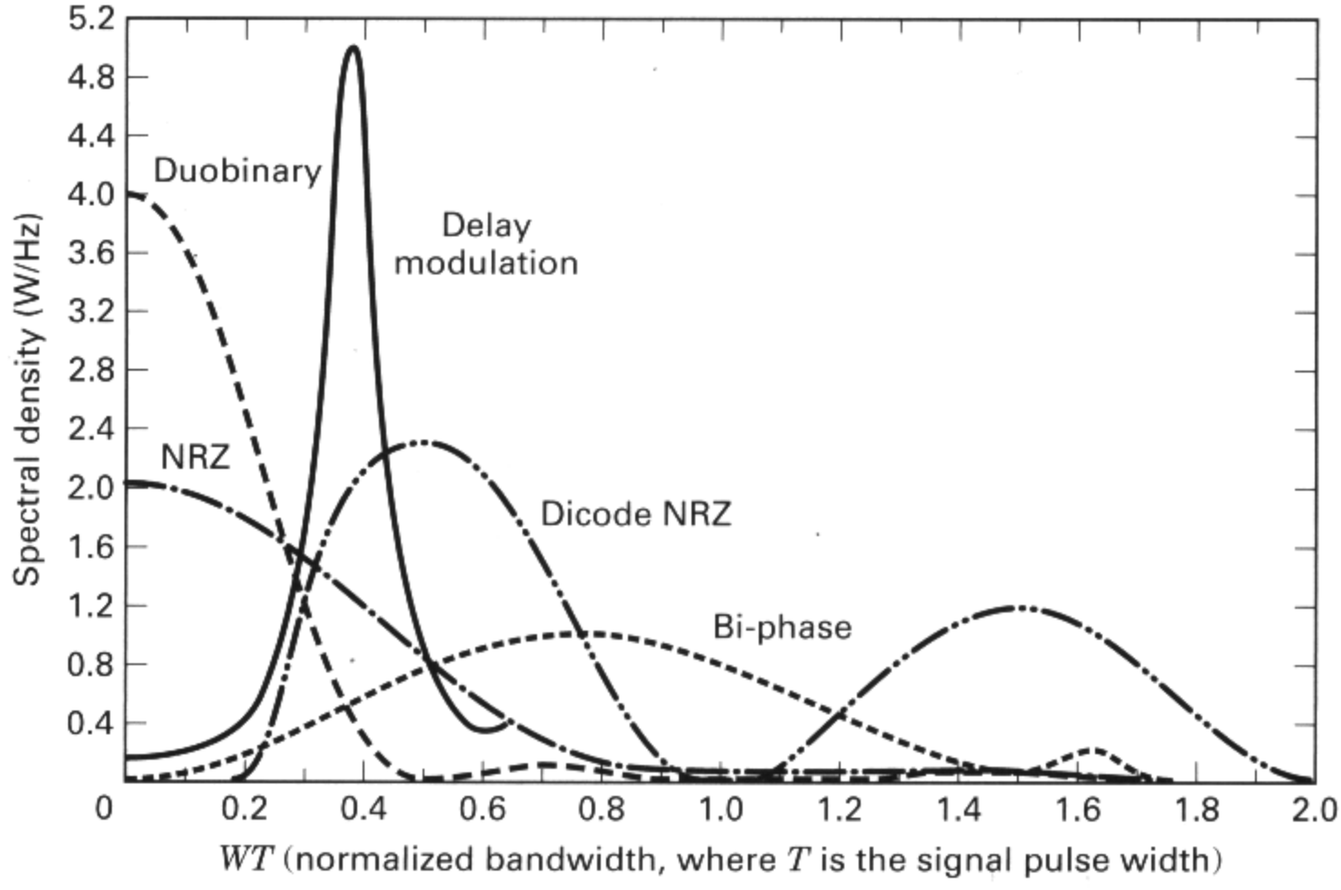


Figure 2.23 Spectral densities of various PCM waveforms.

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.)

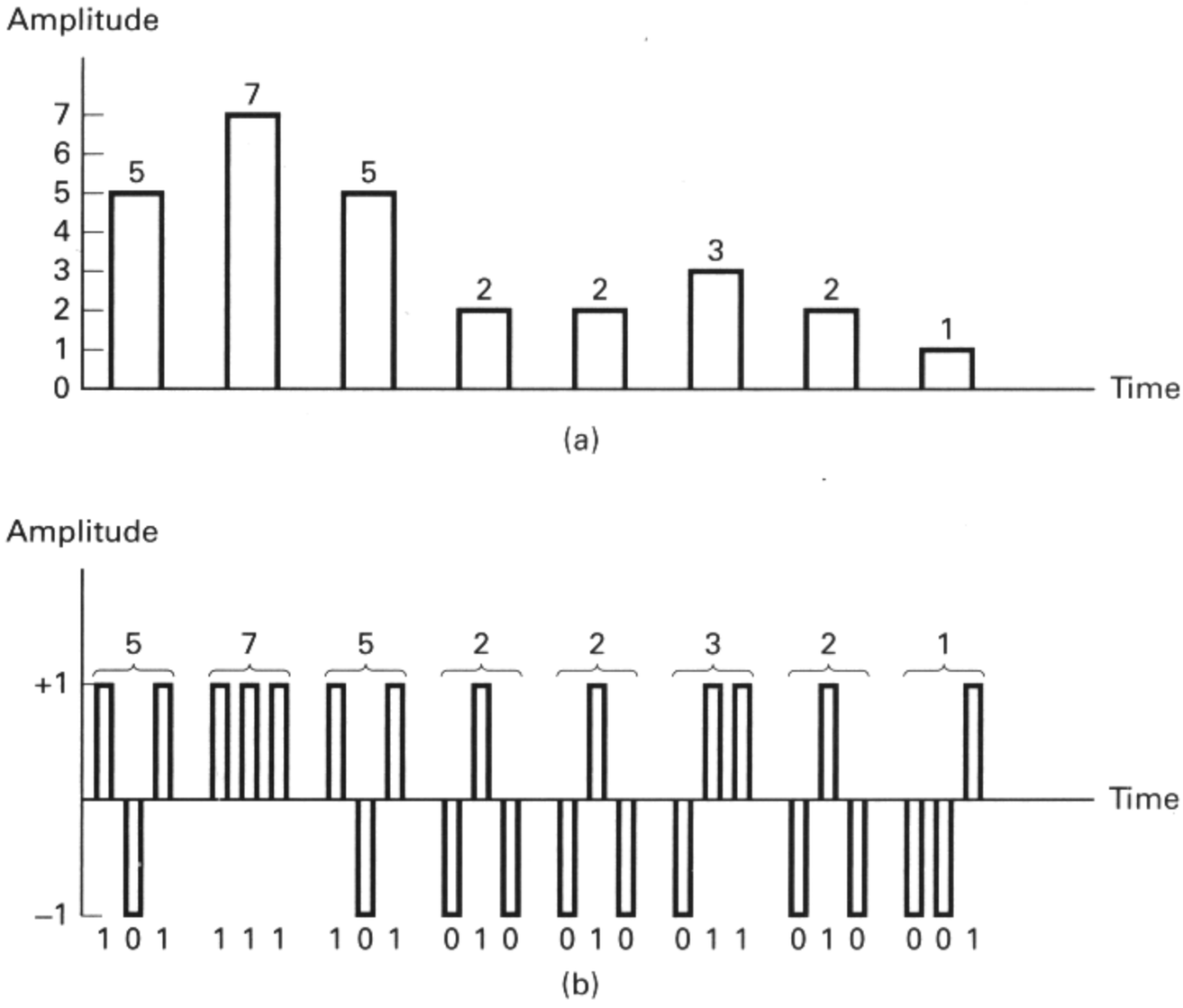


Figure 2.24 Pulse code modulation signaling. (a) Eight-level signaling. (b) Two-level signaling.

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 $\pm 1\%$ 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

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

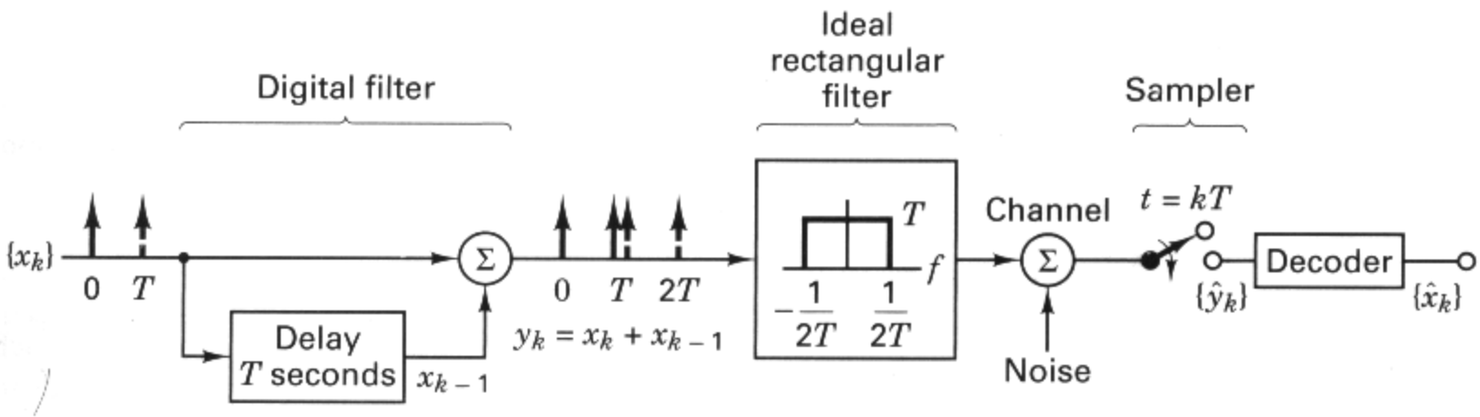


Figure 2.25 Duobinary signaling.

Hence, the $\{y_k\}$ amplitudes are not independent; each y_k digit carries with it the *memory* of the prior digit. The ISI introduced to each y_k digit comes only from the preceding x_{k-1} digit. This correlation between the pulse amplitudes of $\{y_k\}$ can be thought of as the controlled ISI introduced by the duobinary coding. Controlled interference is the essence of this novel technique because at the detector, such controlled interference can be removed as easily as it was added. The $\{y_k\}$ sequence is followed by the ideal Nyquist filter that does not introduce any ISI. In Figure 2.25, at the receiver sampler, we would expect to recover the sequence $\{y_k\}$ exactly in the absence of noise. Since all systems experience noise contamination, we shall refer to the *received* $\{y_k\}$ as the estimate of $\{y_k\}$ and denote it $\{\hat{y}_k\}$. Removing the controlled interference with the duobinary decoder yields an estimate of $\{x_k\}$ which we shall denote as $\{\hat{x}_k\}$.

2.9.2 Duobinary Decoding

If the binary digit x_k is equal to ± 1, then using Equation (2.29), y_k has one of three possible values: + 2, 0, or –2. The duobinary code results in a three-level output: in general, for M-ary transmission, partial response signaling results in $2M - 1$ output levels. The decoding procedure involves the inverse of the coding procedure, namely, subtracting the x_{k-1} decision from the y_k digit. Consider the following coding/decoding example.

Example 2.4 Duobinary Coding and Decoding

Use Equation (2.29) to demonstrate duobinary coding and decoding for the following sequence: $\{x_k\} = 0\ 0\ 1\ 0\ 1\ 1\ 0$. Consider the first bit of the sequence to be a startup digit, not part of the data.

Solution

Binary digit sequence $\{x_k\}$:	0	0	1	0	1	1	0
Bipolar amplitudes $\{x_k\}$:	–1	–1	+1	–1	+1	+1	–1
Coding rule: $y_k = x_k + x_{k-1}$:		–2	0	0	0	2	0

Decoding decision rule:

If $\hat{y}_k = 2$, decide that $\hat{x}_k = +1$ (or binary one).

If $\hat{y}_k = -2$, decide that $\hat{x}_k = -1$ (or binary zero).

If $\hat{y}_k = 0$, decide opposite of the previous decision.

Decoded bipolar sequence $\{\hat{x}_k\}$:	−1	+1	−1	+1	+1	−1
Decoded binary sequence $\{\hat{x}_k\}$:	0	1	0	1	1	0

The decision rule simply implements the subtraction of each $\hat{x}_{k-1}$ decision from each $\hat{y}_k$. One drawback of this detection technique is that once an error is made, it tends to propagate, causing further errors, since present decisions depend on prior decisions. A means of avoiding this error propagation is known as *precoding*.

2.9.3 Precoding

Precoding is accomplished by first differentially encoding the $\{x_k\}$ binary sequence into a new $\{w_k\}$ binary sequence by means of the equation:

$$w_k = x_k \oplus w_{k-1} \tag{2.30}$$

where the symbol $\oplus$ represents modulo-2 addition (equivalent to the logical *exclusive-or* operation) of the binary digits. The rules of modulo-2 addition are as follows:

$$0 \oplus 0 = 0$$
$$0 \oplus 1 = 1$$
$$1 \oplus 0 = 1$$
$$1 \oplus 1 = 0$$

The $\{w_k\}$ binary sequence is then converted to a bipolar pulse sequence, and the coding operation proceeds in the same way as it did in Example 2.4. However, with precoding, the detection process is quite different from the detection of ordinary duobinary, as shown below in Example 2.5: The precoding model is shown in Figure 2.26; in this figure it is implicit that the modulo-2 addition producing the precoded $\{w_k\}$ sequence is performed on the *binary* digits, while the digital filtering producing the $\{y_k\}$ sequence is performed on the *bipolar* pulses.

Example 2.5 Duobinary Precoding

Illustrate the duobinary coding and decoding rules when using the differential precoding of Equation (2.30). Assume the same $\{x_k\}$ sequence as that given in Example 2.4.

Solution

Binary digit sequence $\{x_k\}$:	0	0	1	0	1	1	0
Precoded sequence $w_k = x_k \oplus w_{k-1}$:	0	0	1	1	0	1	1
Bipolar sequence $\{w_k\}$:	−1	−1	+1	+1	−1	+1	+1
Coding rule: $y_k = w_k + w_{k-1}$:		−2	0	+2	0	0	+2

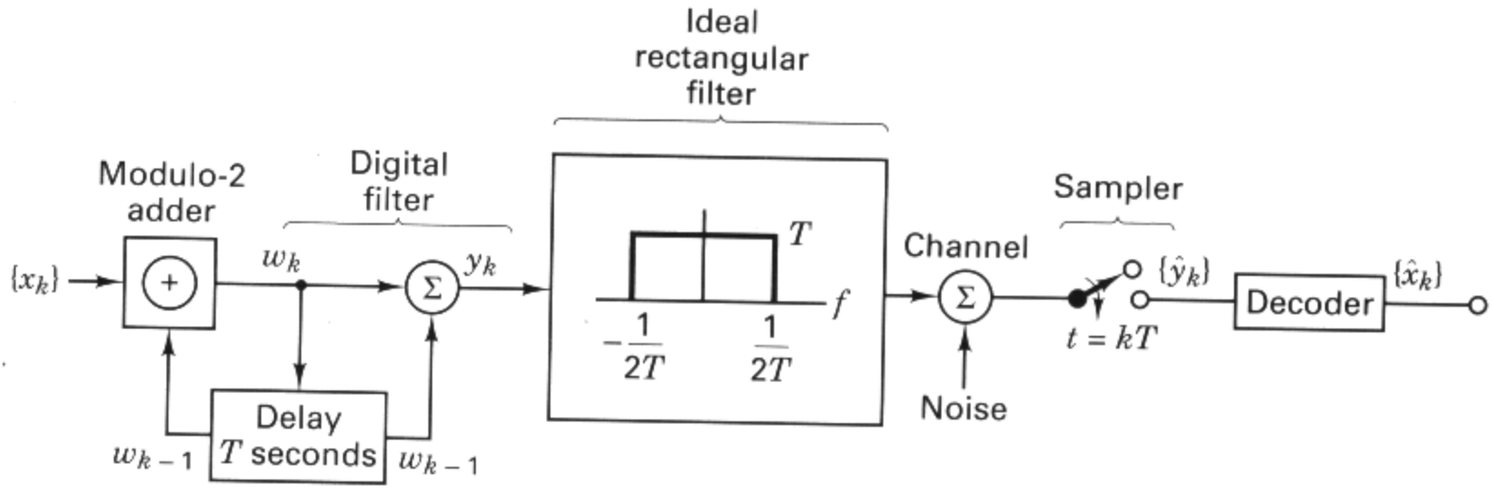


Figure 2.26 Precoded duobinary signaling.

Decoding decision rule: If $\hat{y}_k = \pm 2$, decide that $\hat{x}_k$ = binary zero.
If $\hat{y}_k = 0$, decide that $\hat{x}_k$ = binary one.

Decoded binary sequence $\{x_k\}$: 0 1 0 1 1 0

The differential precoding enables us to decode the $\{\hat{y}_k\}$ sequence by making a decision on each received sample singly, without resorting to prior decisions that could be in error. The major advantage is that in the event of a digit error due to noise, such an error does not propagate to other digits. Notice that the first bit in the differentially precoded binary sequence $\{w_k\}$ is an arbitrary choice. If the startup bit in $\{w_k\}$ had been chosen to be a binary one instead of a binary zero, the decoded result would have been the same.

2.9.4 Duobinary Equivalent Transfer Function

In Section 2.9.1, we described the duobinary transfer function as a digital filter incorporating a one-digit delay followed by an ideal rectangular transfer function. Let us now examine an equivalent model. The Fourier transform of a delay can be described as $e^{-j2\pi fT}$ (see Section A.3.1); therefore, the input digital filter of Figure 2.25 can be characterized as the frequency transfer function

$$H_1(f) = 1 + e^{-j2\pi fT} \tag{2.31}$$

The transfer function of the ideal rectangular filter, is

$$H_2(f) = \begin{cases} T & \text{for } |f| < \dfrac{1}{2T} \\ 0 & \text{elsewhere} \end{cases} \tag{2.32}$$

The overall equivalent transfer function of the digital filter cascaded with the ideal rectangular filter is then given by

$$
\begin{aligned}
H_e(f) &= H_1(f)H_2(f) \qquad \text{for } |f| < \frac{1}{2T} \\
&= (1 + e^{-j2\pi fT})T \\
&= T(e^{j\pi fT} + e^{-j\pi fT})e^{-j\pi fT}
\end{aligned} \tag{2.33}
$$

so that

$$
|H_e(f)| = \begin{cases} 2T\cos \pi fT & \text{for } |f| < \dfrac{1}{2T} \\ 0 & \text{elsewhere} \end{cases} \tag{2.34}
$$

Thus $H_e(f)$, the composite transfer function for the cascaded digital and rectangular filters, has a gradual roll-off to the band edge, as can be seen in Figure 2.27a. The transfer function can be approximated by using realizable analog filtering; a separate digital filter is not needed. The duobinary equivalent $H_e(f)$ is called a *cosine filter* [8]. The cosine filter should not be confused with the *raised* cosine filter (described in Chapter 3, Section 3.3.1). The corresponding impulse response $h_e(t)$, found by taking the inverse Fourier transform of $H_e(f)$ in Equation (2.33) is

$$
h_e(t) = \text{sinc}\left(\frac{t}{T}\right) + \text{sinc}\left(\frac{t - T}{T}\right) \tag{2.35}
$$

and is plotted in Figure 2.27b. For every impulse $\delta(t)$ at the input of Figure 2.25, the output is $h_e(t)$ with an appropriate polarity. Notice that there are only two nonzero samples at T-second intervals, giving rise to controlled ISI from the adjacent bit. The introduced ISI is eliminated by use of the decoding procedure discussed in Section 2.9.2. Although the cosine filter is noncausal and therefore nonrealizable, it can be easily approximated. The implementation of the precoded duobinary technique described in Section 2.9.3 can be accomplished by first differentially encoding the binary sequence $\{x_k\}$ into the sequence $\{w_k\}$ (see Example 2.5). The pulse sequence $\{w_k\}$ is then filtered by the equivalent cosine characteristic described in Equation (2.34).

2.9.5 Comparison of Binary with Duobinary Signaling

The duobinary technique introduces correlation between pulse amplitudes, whereas the more restrictive Nyquist criterion assumes that the transmitted pulse amplitudes are independent of one another. We have shown that duobinary signaling can exploit this introduced correlation to achieve zero ISI signal transmission, using a smaller system bandwidth than is otherwise possible. Do we get this performance improvement without paying a price? No, such is rarely the case with engineering design options—there is almost always a trade-off involved. We saw that duobinary coding requires three levels, compared with the usual two levels for binary coding. Recall our discussion in Section 2.8.5, where we compared the performance and the required signal power for making eight-level PAM decisions versus two-level (PCM) decisions. For a fixed amount of signal power, the ease of making reliable decisions is inversely related to the number of levels that must be

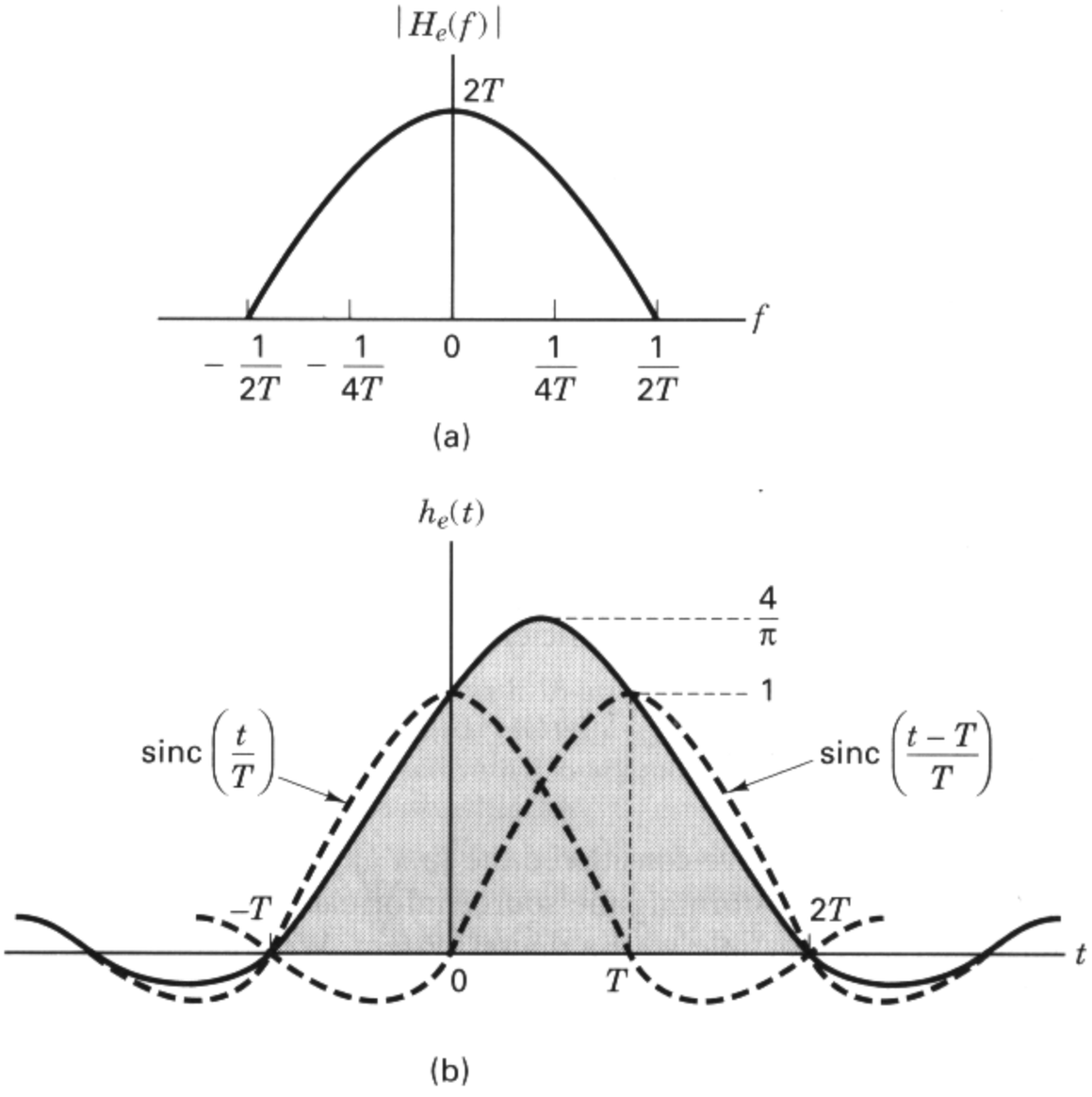


Figure 2.27 Duobinary transfer function and pulse shape. (a) Cosine filter. (b) Impulse response of the cosine filter.

distinguished in each waveform. Therefore, it should be no surprise that although duobinary signaling accomplishes the zero ISI requirement with minimum bandwidth, duobinary signaling also requires more power than binary signaling, for equivalent performance against noise. For a given probability of bit error (P_B), duobinary signaling requires approximately 2.5 dB greater SNR than binary signaling, while using only $1/(1 + r)$ the bandwidth that binary signaling requires [7], where r is the filter roll-off.

2.9.6 Polybinary Signaling

Duobinary signaling can be extended to more than three digits or levels, resulting in greater bandwidth efficiency; such systems are called *polybinary* [7, 9]. Consider that a binary message with two signaling levels is transformed into a signal with j signaling levels numbered consecutively from zero to $(j - 1)$. The transformation from binary to polybinary takes place in two steps. First, the original sequence $\{x_k\}$, consisting of binary ones and zeros, is converted into another binary sequence $\{y_k\}$,

as follows. The present binary digit of sequence $\{y_k\}$ is formed from the modulo-2 addition of the $(j-2)$ immediately preceding digits of sequence $\{y_k\}$ and the present digit x_k. For example, let

$$y_k = x_k \oplus y_{k-1} \oplus y_{k-2} \oplus y_{k-3} \tag{2.36}$$

Here x_k represents the input binary digit and y_k the kth encoded digit. Since the expression involves $(j-2) = 3$ bits preceding y_k, there are $j = 5$ signaling levels. Next, the binary sequence $\{y_k\}$ is transformed into a polybinary pulse train $\{z_k\}$ by adding *algebraically* the present bit of sequence $\{y_k\}$ to the $(j-2)$ preceding bits of $\{y_k\}$. Therefore, z_k modulo-2 = x_k, and the binary elements one and zero are mapped into even- and odd-valued pulses in the sequence $\{z_k\}$. Note that each digit in $\{z_k\}$ can be independently detected despite the strong correlation between bits. The primary advantage of such a signaling scheme is the redistribution of the spectral density of the original sequence $\{x_k\}$, so as to favor the low frequencies, thus improving system bandwidth efficiency.

2.10 CONCLUSION

In this chapter we have considered the first important step in any digital communication system, transforming the source information (both textual and analog) to a form that is compatible with a digital system. We treated various aspects of sampling, quantization (both uniform and nonuniform), and pulse code modulation (PCM). We considered the selection of pulse waveforms for the transmission of baseband signals through the channel. We also introduced the duobinary concept of adding a controlled amount of ISI to achieve an improvement in bandwidth efficiency at the expense of an increase in power.

REFERENCES

1. Black, H. S., *Modulation Theory,* D. Van Nostrand Company, Princeton, N.J., 1953.
2. Oppenheim, A. V., *Applications of Digital Signal Processing,* Prentice-Hall, Inc., Englewood Cliffs, N.J., 1978.
3. Stiltz, H., ed., *Aerospace Telemetry,* Vol. 1, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1961, p. 179.
4. Hecht, M., and Guida, A., "Delay Modulation," *Proc. IEEE,* vol. 57, no. 7, July 1969, pp. 1314–1316.
5. Deffebach, H. L., and Frost, W. O., "A Survey of Digital Baseband Signaling Techniques," *NASA Technical Memorandum NASATM X-64615,* June 30, 1971.
6. Lender, A., "The Duobinary Technique for High Speed Data Transmission," *IEEE Trans. Commun. Electron.,* vol. 82, May 1963, pp. 214–218.
7. Lender, A., "Correlative (Partial Response) Techniques and Applications to Digital Radio Systems," in K. Feher, *Digital Communications: Microwave Applications,* Prentice-Hall, Inc., Englewood Cliffs, N.J., 1981, Chap. 7.

8. Couch, L. W., II, *Digital and Analog Communication Systems,* Macmillan Publishing Company, New York, 1982.
9. Lender, A., "Correlative Digital Communication Techniques," *IEEE Trans. Commun. Technol.,* Dec. 1964, pp. 128–135.

PROBLEMS

2.1. You want to transmit the word "HOW" using an 8-ary system.
(a) Encode the word "HOW" into a sequence of bits, using 7-bit ASCII coding, followed by an eighth bit for error detection, per character. The eighth bit is chosen so that the number of ones in the 8 bits is an even number. How many total bits are there in the message?
(b) Partition the bit stream into $k = 3$ bit segments. Represent each of the 3-bit segments as an octal number (symbol). How many octal symbols are there in the message?
(c) If the system were designed with 16-ary modulation, how many symbols would be used to represent the word "HOW"?
(d) If the system were designed with 256-ary modulation, how many symbols would be used to represent the word "HOW"?

2.2. We want to transmit 800 characters/s, where each character is represented by its 7-bit ASCII codeword, followed by an eighth bit for error detection, per character, as in Problem 2.1. A multilevel PAM waveform with $M = 16$ levels is used.
(a) What is the effective transmitted bit rate?
(b) What is the symbol rate?

2.3. We wish to transmit a 100-character alphanumeric message in 2 s, using 7-bit ASCII coding, followed by an eighth bit for error detection, per character, as in Problem 2.1. A multilevel PAM waveform with $M = 32$ levels is used.
(a) Calculate the effective transmitted bit rate and the symbol rate.
(b) Repeat part (a) for 16-level PAM, eight-level PAM, four-level PAM, and PCM (binary) waveforms.

2.4. Given an analog waveform that has been sampled at its Nyquist rate, f_s, using natural sampling, prove that a waveform (proportional to the original waveform) can be recovered from the samples, using the recovery techniques shown in Figure P2.1. The parameter mf_s is the frequency of the local oscillator, where m is an integer.

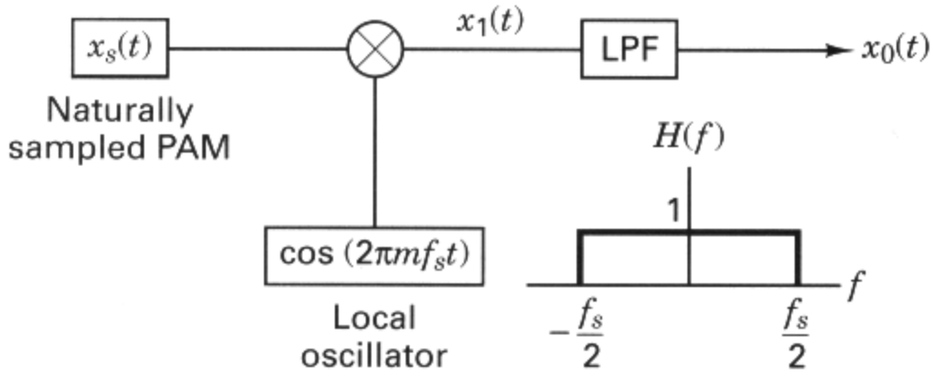


Figure P2.1

2.5. An analog signal is sampled at its Nyquist rate $1/T_s$, and quantized using L quantization levels. The derived digital signal is then transmitted on some channel.

(a) Show that the time duration, T, of one bit of the transmitted binary encoded signal must satisfy $T \leq T_s/(\log_2 L)$.

(b) When is the equality sign valid?

2.6. Determine the number of quantization levels that are implied if the number of bits per sample in a given PCM code is (a) 5; (b) 8; (c) x.

2.7. Determine the minimum sampling rate necessary to sample and perfectly reconstruct the signal $x(t) = \sin(6280t)/(6280t)$.

2.8. Consider an audio signal with spectral components limited to the frequency band 300 to 3300 Hz. Assume that a sampling rate of 8000 samples/s will be used to generate a PCM signal. Assume that the ratio of peak signal power to average quantization noise power at the output needs to be 30 dB.

(a) What is the minimum number of uniform quantization levels needed, and what is the minimum number of bits per sample needed?

(b) Calculate the system bandwidth (as specified by the main spectral lobe of the signal) required for the detection of such a PCM signal.

2.9. A waveform, $x(t) = 10\cos(1000t + \pi/3) + 20\cos(2000t + \pi/6)$ is to be uniformly sampled for digital transmission.

(a) What is the maximum allowable time interval between sample values that will ensure perfect signal reproduction?

(b) If we want to reproduce 1 hour of this waveform, how many sample values need to be stored?

2.10. (a) A waveform that is bandlimited to 50 kHz is sampled every 10 μs. Show graphically that these samples uniquely characterize the waveform. (Use a sinusoidal example for simplicity. Avoid sampling at points where the waveform equals zero.)

(b) If samples are taken 30 μs apart instead of 10 μs, show graphically that waveforms other than the original can be characterized by the samples.

2.11. Use the method of convolution to illustrate the effect of undersampling the waveform $x(t) = \cos 2\pi f_0 t$ for a sampling rate of $f_s = \frac{3}{2} f_0$.

2.12. Aliasing will not occur if the sampling rate is greater than twice the signal bandwidth. However, perfectly bandlimited signals do not occur in nature. Hence, there is always some aliasing present.

(a) Suppose that a filtered signal has a spectrum described by a Butterworth filter with order $n = 6$, and upper cutoff frequency $f_u = 1000$ Hz. What sampling rate is required so that aliasing is reduced to the −50 dB point in the power spectrum?

(b) Repeat for a Butterworth filter with order n =12.

213. (a) Sketch the complete $\mu = 10$ compression characteristic that will handle input voltages in the range −5 to +5 V.

(b) Plot the corresponding expansion characteristic.

(c) Draw a 16-level nonuniform quantizer characteristic that corresponds to the $\mu = 10$ compression characteristic.

2.14. The information in an analog waveform, whose maximum frequency $f_m = 4000$ Hz, is to be transmitted using a 16-level PAM system. The quantization distortion must not exceed ± 1% of the peak-to-peak analog signal.

(a) What is the minimum number of bits per sample or bits per PCM word that should be used in this PAM transmission system?

(b) What is the minimum required sampling rate, and what is the resulting bit rate?

(c) What is the 16-ary PAM symbol transmission rate?

2.15. A signal in the frequency range 300 to 3300 Hz is limited to a peak-to-peak swing of 10 V. It is sampled at 8000 samples/s and the samples are quantized to 64 evenly spaced levels. Calculate and compare the bandwidths and ratio of peak signal power to rms quantization noise if the quantized samples are transmitted either as binary pulses or as four-level pulses. Assume that the system bandwidth is defined by the main spectral lobe of the signal.

2.16. In the compact disc (CD) digital audio system, an analog signal is digitized so that the ratio of the peak-signal power to the peak-quantization noise power is at least 96 dB. The sampling rate is 44.1 kilosamples/s.

(a) How many quantization levels of the analog signal are needed for $(S/N_q)_{\text{peak}} = 96$ dB?

(b) How many bits per sample are needed for the number of levels found in part (a)?

(c) What is the data rate in bits/s?

2.17. Calculate the difference in required signal power between two PCM waveforms, NRZ and RZ, assuming that each signaling scheme has the same requirements for data-rate and bit-error probability. Also assume equally likely signaling, and that the difference between the high-voltage and low-voltage levels is the same for both the NRZ and RZ schemes. If there is a power advantage in using one of the signaling schemes, what, if any, is the disadvantage in using it?

2.18. In the year 1962, AT&T first offered digital telephone transmission referred to as T1 service. With this service, each T1 frame is partitioned into 24 channels or time slots. Each time slot contains 8 bits (one speech sample), and there is one additional bit per frame for alignment. The frame is sampled at the Nyquist rate of 8000 samples/s, and the bandwidth used for transmitting the composite signal is 386 kHz. Find the bandwidth efficiency (bits/s/Hz) for this signaling scheme.

2.19. **(a)** Consider that you desire a digital transmission system, such that the quantization distortion of any audio source does not exceed ± 2% of the peak-to-peak analog signal voltage. If the audio signal bandwidth and the allowable transmission bandwidth are each 4000 Hz, and sampling takes place at the Nyquist rate, what value of bandwidth efficiency (bits/s/Hz) is required?

(b) Repeat part (a) except that the audio signal bandwidth is 20 kHz (high fidelity), yet the available transmission bandwidth is still 4000 Hz.

QUESTIONS

2.1 What are the similarities and differences between the terms "formatting" and "source coding"? (See Chapter 2, introduction.)

2.2 In the process of *formatting* information, why is it often desirable to perform *oversampling?* (See Section 2.4.3.)

2.3 In using pulse code modulation (PCM) for digitizing analog information, explain how the parameters *fidelity, bandwidth,* and *time delay* can be traded off. (See Section 2.6.)

2.4 Why is it often preferred to use units of normalized bandwidth, WT (or time-bandwidth product), compared with bandwidth alone? (See Section 2.8.3.)

EXERCISES

Using the Companion CD, run the exercises associated with Chapter 2.