CHAPTER 3

DATA TRANSMISSION

Toto, I've got a feeling we're not in Kansas anymore.

Judy Garland in *The Wizard of Oz*

KEY POINTS

- All of the forms of information that are discussed in this book (voice, data, image, video) can be represented by electromagnetic signals. Depending on the transmission medium and the communications environment, either analog or digital signals can be used to convey information.
- Any electromagnetic signal, analog or digital, is made up of a number of constituent frequencies. A key parameter that characterizes the signal is bandwidth, which is the width of the range of frequencies that comprises the signal. In general, the greater the bandwidth of the signal, the greater its information-carrying capacity.
- A major problem in designing a communications facility is transmission impairment. The most significant impairments are attenuation, attenuation distortion, delay distortion, and the various types of noise. The various forms of noise include thermal noise, intermodulation noise, crosstalk, and impulse noise. For analog signals, transmission impairments introduce random effects that degrade the quality of the received information and may affect intelligibility. For digital signals, transmission impairments may cause bit errors at the receiver.
- The designer of a communications facility must deal with four factors: the bandwidth of the signal, the data rate that is used for digital information, the amount of noise and other impairments, and the level of error rate that is acceptable. The bandwidth is limited by the transmission medium and the desire to avoid interference with other nearby signals. Because bandwidth is a scarce resource, we would like to maximize the data rate that is achieved in a given bandwidth. The data rate is limited by the bandwidth, the presence of impairments, and the error rate that is acceptable.

The successful transmission of data depends principally on two factors: the quality of the signal being transmitted and the characteristics of the transmission medium. The objective of this chapter and the next is to provide the reader with an intuitive feeling for the nature of these two factors.

The first section presents some concepts and terms from the field of electrical engineering. This should provide sufficient background to deal with the remainder of the chapter. Section 3.2 clarifies the use of the terms *analog* and *digital*. Either analog or digital data may be transmitted using either analog or digital signals. Furthermore, it is common for intermediate processing to be performed between source and destination, and this processing has either an analog or digital character.

Section 3.3 looks at the various impairments that may introduce errors into the data during transmission. The chief impairments are attenuation, attenuation distortion, delay distortion, and the various forms of noise. Finally, we look at the important concept of channel capacity.

3.1 CONCEPTS AND TERMINOLOGY

In this section we introduce some concepts and terms that will be referred to throughout the rest of the chapter and, indeed, throughout Part Two.

Transmission Terminology

Data transmission occurs between transmitter and receiver over some transmission medium. Transmission media may be classified as guided or unguided. In both cases, communication is in the form of electromagnetic waves. With **guided media**, the waves are guided along a physical path; examples of guided media are twisted pair, coaxial cable, and optical fiber. **Unguided media**, also called **wireless**, provide a means for transmitting electromagnetic waves but do not guide them; examples are propagation through air, vacuum, and seawater.

The term **direct link** is used to refer to the transmission path between two devices in which signals propagate directly from transmitter to receiver with no intermediate devices, other than amplifiers or repeaters used to increase signal strength. Note that this term can apply to both guided and unguided media.

A guided transmission medium is **point to point** if it provides a direct link between two devices and those are the only two devices sharing the medium. In a **multipoint** guided configuration, more than two devices share the same medium.

A transmission may be simplex, half duplex, or full duplex. In **simplex** transmission, signals are transmitted in only one direction; one station is transmitter and the other is receiver. In **half-duplex** operation, both stations may transmit, but only one at a time. In **full-duplex** operation, both stations may transmit simultaneously. In the latter case, the medium is carrying signals in both directions at the same time. How this can be is explained in due course. We should note that the definitions just given are the ones in common use in the United States (ANSI definitions). Elsewhere (ITU-T definitions), the term *simplex* is used to correspond to *half duplex* as defined previously, and *duplex* is used to correspond to *full duplex* as just defined.

Frequency, Spectrum, and Bandwidth

In this book, we are concerned with electromagnetic signals used as a means to transmit data. At point 3 in Figure 1.3, a signal is generated by the transmitter and transmitted over a medium. The signal is a function of time, but it can also be expressed as a function of frequency; that is, the signal consists of components of different frequencies. It turns out that the **frequency domain** view of a signal is more important to an understanding of data transmission than a **time domain** view. Both views are introduced here.

Time Domain Concepts Viewed as a function of time, an electromagnetic signal can be either analog or digital. An **analog signal** is one in which the signal intensity

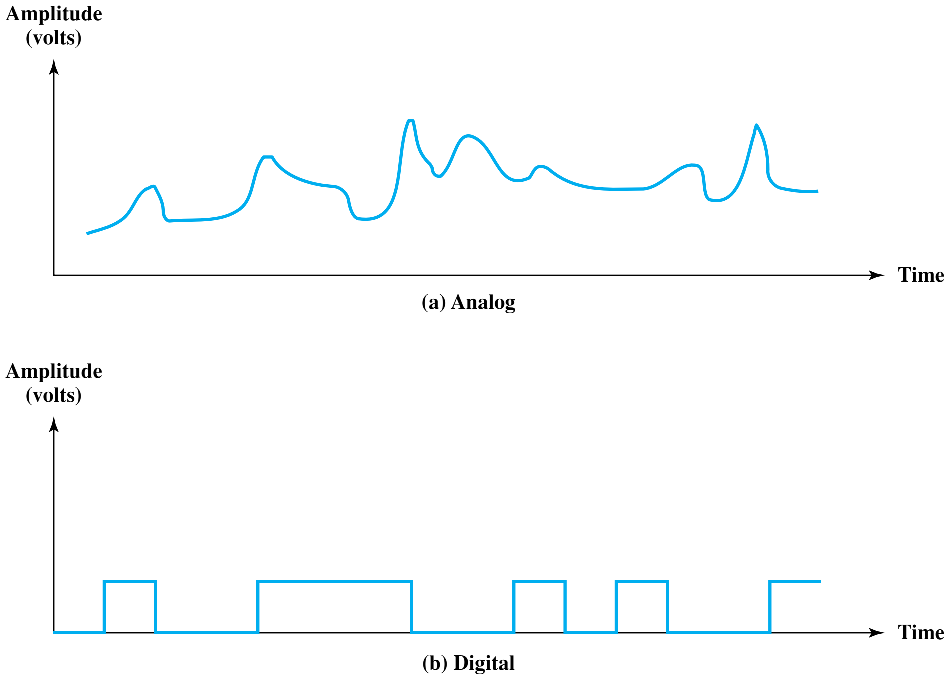


Figure 3.1 Analog and Digital Waveforms

varies in a smooth fashion over time. In other words, there are no breaks or discontinuities in the signal.[1] A **digital signal** is one in which the signal intensity maintains a constant level for some period of time and then abruptly changes to another constant level.[2] Figure 3.1 shows an example of each kind of signal. The continuous signal might represent speech, and the discrete signal might represent binary 1s and 0s.

The simplest sort of signal is a **periodic signal**, in which the same signal pattern repeats over time. Figure 3.2 shows an example of a periodic continuous signal (sine wave) and a periodic discrete signal (square wave). Mathematically, a signal $s(t)$ is defined to be periodic if and only if

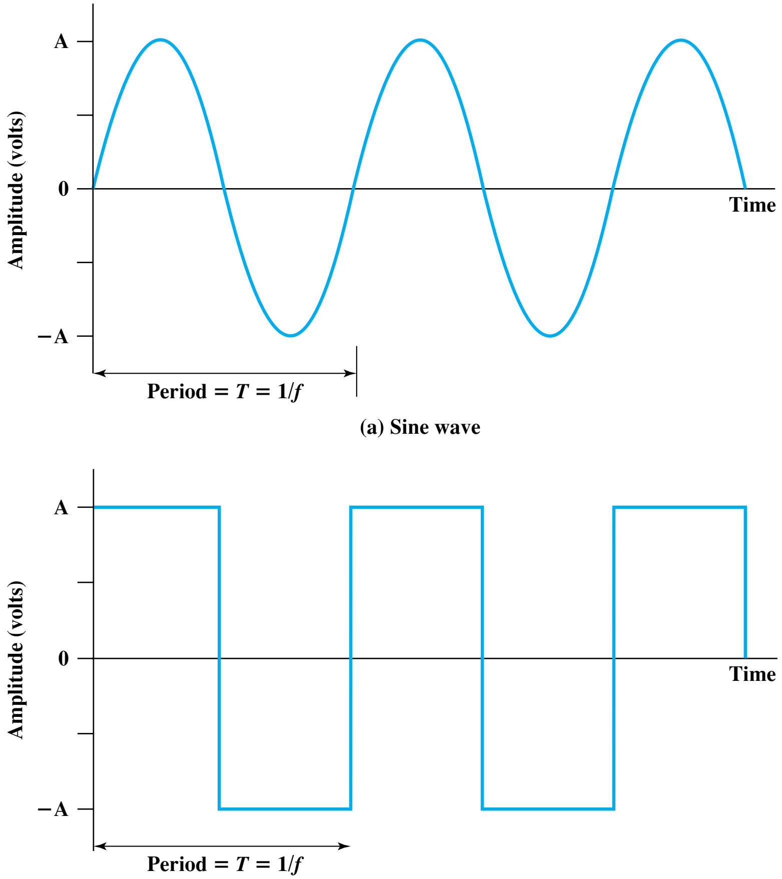
$$s(t + T) = s(t) \qquad -\infty < t < +\infty$$

where the constant T is the period of the signal (T is the smallest value that satisfies the equation). Otherwise, a signal is **aperiodic**.

The sine wave is the fundamental periodic signal. A general sine wave can be represented by three parameters: peak amplitude (A), frequency (f), and phase (ϕ). The **peak amplitude** is the maximum value or strength of the signal over time; typically, this value is measured in volts. The **frequency** is the rate [in cycles per

[1] A mathematical definition: a signal $s(t)$ is continuous if $\lim_{t \to a} s(t) = s(a)$ for all a.

[2] This is an idealized definition. In fact, the transition from one voltage level to another will not be instantaneous, but there will be a small transition period. Nevertheless, an actual digital signal approximates closely the ideal model of constant voltage levels with instantaneous transitions.



(a) Sine wave

(b) Square wave

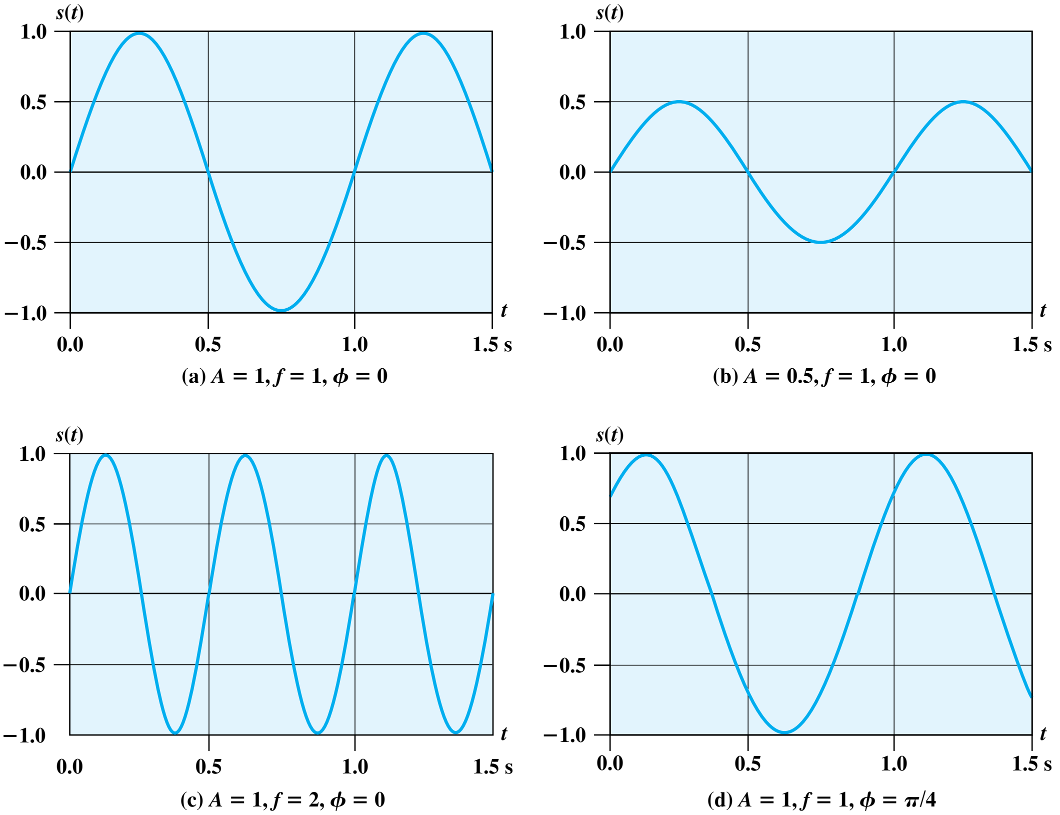
Figure 3.2 Examples of Periodic Signals

second, or Hertz (Hz)] at which the signal repeats. An equivalent parameter is the **period** (T) of a signal, which is the amount of time it takes for one repetition; therefore, $T = 1/f$. **Phase** is a measure of the relative position in time within a single period of a signal, as is illustrated subsequently. More formally, for a periodic signal $f(t)$, phase is the fractional part t/T of the period T through which t has advanced relative to an arbitrary origin. The origin is usually taken as the last previous passage through zero from the negative to the positive direction.

The general sine wave can be written

$$s(t) = A \sin(2\pi f t + \phi)$$

A function with the form of the preceding equation is known as a **sinusoid**. Figure 3.3 shows the effect of varying each of the three parameters. In part (a) of the figure, the frequency is 1 Hz; thus the period is $T = 1$ second. Part (b) has the same frequency and phase but a peak amplitude of 0.5. In part (c) we have $f = 2$, which is equivalent to $T = 0.5$. Finally, part (d) shows the effect of a phase shift of $\pi/4$ radians, which is 45 degrees (2π radians $= 360° = 1$ period).



(a) $A = 1, f = 1, \phi = 0$

(b) $A = 0.5, f = 1, \phi = 0$

(c) $A = 1, f = 2, \phi = 0$

(d) $A = 1, f = 1, \phi = \pi/4$

Figure 3.3 $s(t) = A \sin(2\pi f t + \phi)$

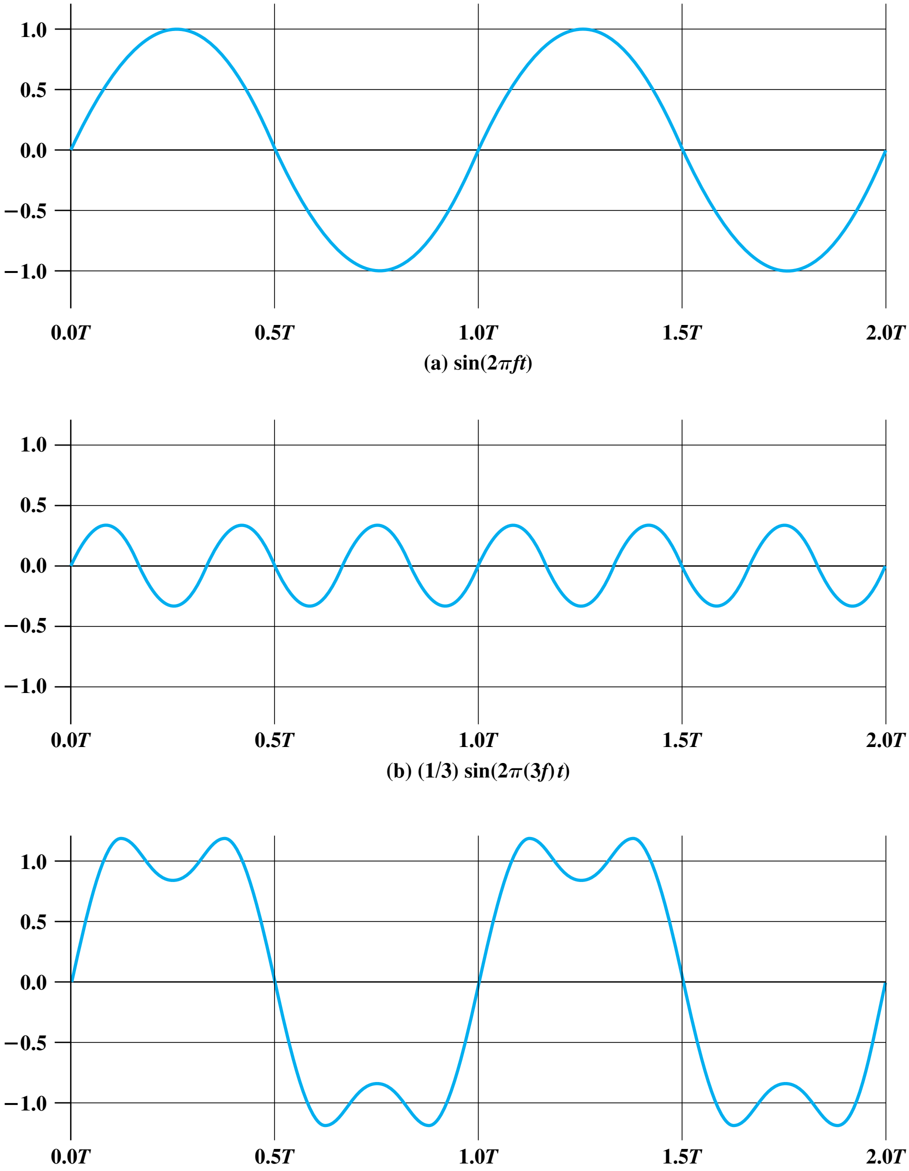
In Figure 3.3, the horizontal axis is time; the graphs display the value of a signal at a given point in space as a function of time. These same graphs, with a change of scale, can apply with horizontal axes in space. In this case, the graphs display the value of a signal at a given point in time as a function of distance. For example, for a sinusoidal transmission (e.g., an electromagnetic radio wave some distance from a radio antenna, or sound some distance from a loudspeaker), at a particular instant of time, the intensity of the signal varies in a sinusoidal way as a function of distance from the source.

There is a simple relationship between the two sine waves, one in time and one in space. The **wavelength** (λ) of a signal is the distance occupied by a single cycle, or, put another way, the distance between two points of corresponding phase of two consecutive cycles. Assume that the signal is traveling with a velocity v. Then the wavelength is related to the period as follows: $\lambda = vT$. Equivalently, $\lambda f = v$. Of particular relevance to this discussion is the case where $v =$ c, the speed of light in free space, which is approximately 3×10^8 m/s.

Frequency Domain Concepts In practice, an electromagnetic signal will be made up of many frequencies. For example, the signal

$$s(t) = [4/\pi) \times (\sin(2\pi f t) + (1/3)\sin(2\pi(3f)t)]$$

is shown in Figure 3.4c. The components of this signal are just sine waves of frequencies f and $3f$; parts (a) and (b) of the figure show these individual components.[3] There are two interesting points that can be made about this figure:



(a) sin(2πft)

(b) (1/3) sin(2π(3f) t)

(c) (4/π) [sin(2πft) + (1/3)sin(2π(3f) t)]

Figure 3.4 Addition of Frequency Components ($T = 1/f$)

[3] The scaling factor of $4/\pi$ is used to produce a wave whose peak amplitude is close to 1.

- The second frequency is an integer multiple of the first frequency. When all of the frequency components of a signal are integer multiples of one frequency, the latter frequency is referred to as the **fundamental frequency**.
- The period of the total signal is equal to the period of the fundamental frequency. The period of the component sin $(2\pi ft)$ is $T = 1/f$, and the period of $s(t)$ is also T, as can be seen from Figure 3.4c.

It can be shown, using a discipline known as Fourier analysis, that any signal is made up of components at various frequencies, in which each component is a sinusoid. By adding together enough sinusoidal signals, each with the appropriate amplitude, frequency, and phase, any electromagnetic signal can be constructed. Put another way, any electromagnetic signal can be shown to consist of a collection of periodic analog signals (sine waves) at different amplitudes, frequencies, and phases. The importance of being able to look at a signal from the frequency perspective (frequency domain) rather than a time perspective (time domain) should become clear as the discussion proceeds. For the interested reader, the subject of Fourier analysis is introduced in Appendix A.

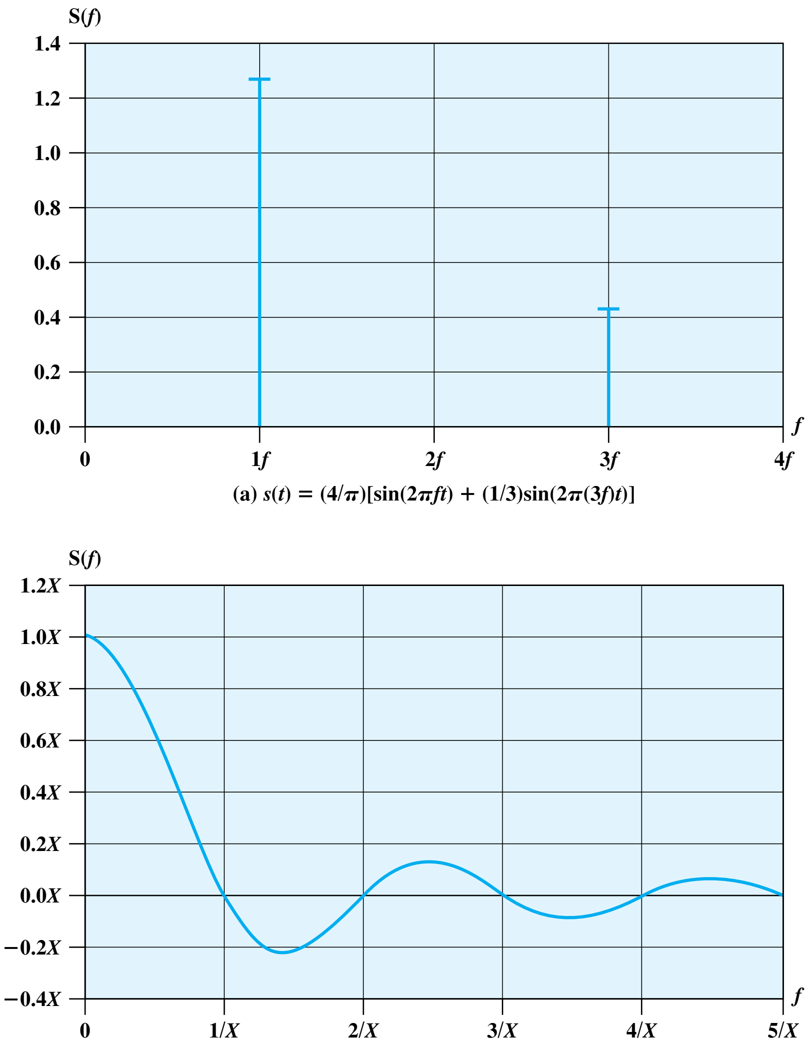
So we can say that for each signal, there is a time domain function $s(t)$ that specifies the amplitude of the signal at each instant in time. Similarly, there is a frequency domain function $S(f)$ that specifies the peak amplitude of the constituent frequencies of the signal. Figure 3.5a shows the frequency domain function for the signal of Figure 3.4c. Note that, in this case, $S(f)$ is discrete. Figure 3.5b shows the frequency domain function for a single square pulse that has the value 1 between $-X/2$ and $X/2$, and is 0 elsewhere.[4] Note that in this case $S(f)$ is continuous and that it has nonzero values indefinitely, although the magnitude of the frequency components rapidly shrinks for larger f. These characteristics are common for real signals.

The **spectrum** of a signal is the range of frequencies that it contains. For the signal of Figure 3.4c, the spectrum extends from f to $3f$. The **absolute bandwidth** of a signal is the width of the spectrum. In the case of Figure 3.4c, the bandwidth is $2f$. Many signals, such as that of Figure 3.5b, have an infinite bandwidth. However, most of the energy in the signal is contained in a relatively narrow band of frequencies. This band is referred to as the **effective bandwidth**, or just **bandwidth**.

One final term to define is **dc component**. If a signal includes a component of zero frequency, that component is a direct current (dc) or constant component. For example, Figure 3.6 shows the result of adding a dc component to the signal of Figure 3.4c. With no dc component, a signal has an average amplitude of zero, as seen in the time domain. With a dc component, it has a frequency term at $f = 0$ and a nonzero average amplitude.

Relationship between Data Rate and Bandwidth We have said that effective bandwidth is the band within which most of the signal energy is concentrated. The meaning of the term *most* in this context is somewhat arbitrary. The important issue

[4] In fact, the function $S(f)$ for this case is symmetric around $f = 0$ and so has values for negative frequencies. The presence of negative frequencies is a mathematical artifact whose explanation is beyond the scope of this book.



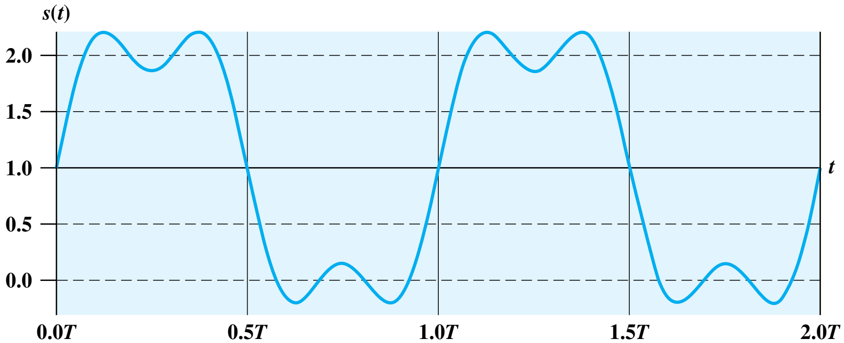
(a) $s(t) = (4/\pi)[\sin(2\pi ft) + (1/3)\sin(2\pi(3f)t)]$

(b) $s(t) = 1 \qquad -X/2 \leq t \leq X/2$

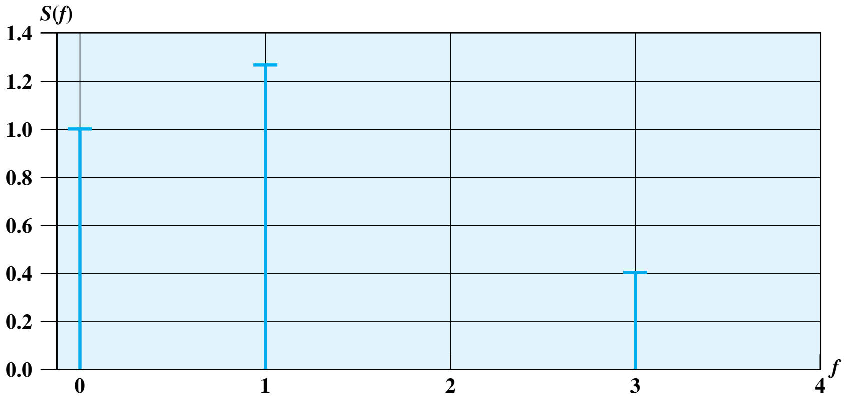
Figure 3.5 Frequency Domain Representations

here is that, although a given waveform may contain frequencies over a very broad range, as a practical matter any transmission system (transmitter plus medium plus receiver) will be able to accommodate only a limited band of frequencies. This, in turn, limits the data rate that can be carried on the transmission medium.

To try to explain these relationships, consider the square wave of Figure 3.2b. Suppose that we let a positive pulse represent binary 0 and a negative pulse represent binary 1. Then the waveform represents the binary stream 0101. . . . The duration of each pulse is $1/(2f)$; thus the data rate is $2f$ bits per second (bps). What are the frequency components of this signal? To answer this question, consider again



(a) $s(t) = 1 + (4/\pi)\,[\sin(2\pi ft) + (1/3)\sin(2\pi(3f)t)]$



(b) $S(f)$

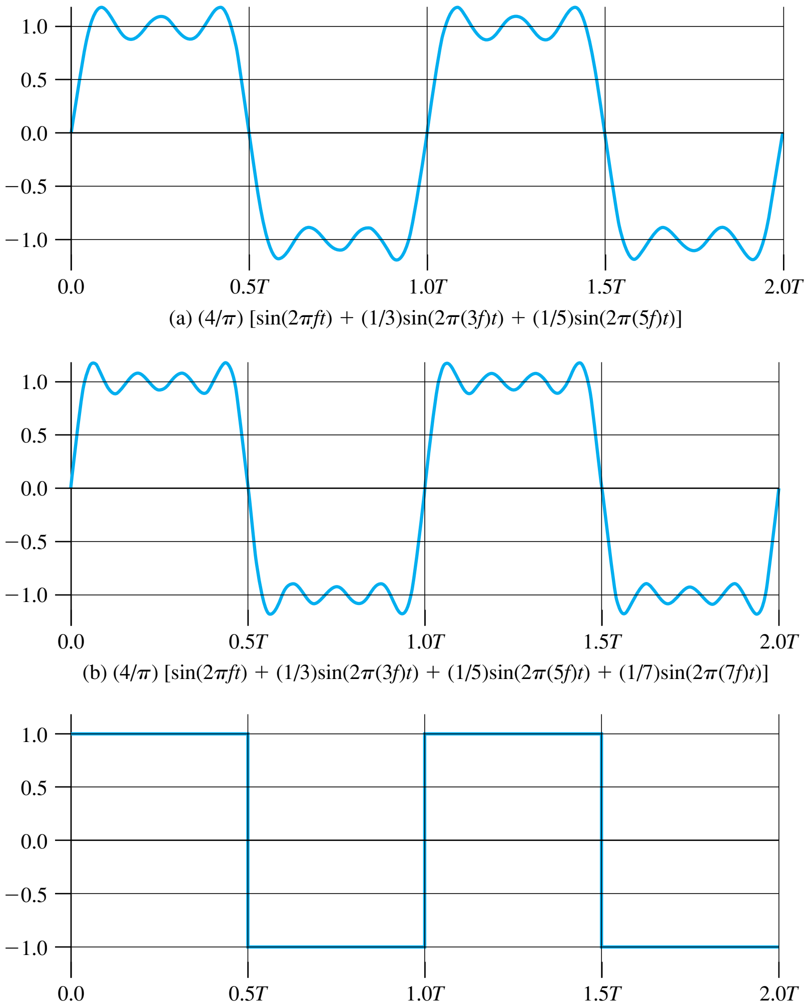
Figure 3.6 Signal with dc Component

Figure 3.4. By adding together sine waves at frequencies *f* and 3*f*, we get a waveform that begins to resemble the original square wave. Let us continue this process by adding a sine wave of frequency 5*f*, as shown in Figure 3.7a, and then adding a sine wave of frequency 7*f*, as shown in Figure 3.7b. As we add additional odd multiples of *f*, suitably scaled, the resulting waveform approaches that of a square wave more and more closely.

Indeed, it can be shown that the frequency components of the square wave with amplitudes A and $-A$ can be expressed as follows:

$$s(t) = A \times \frac{4}{\pi} \times \sum_{k\ \mathrm{odd},k=1}^{\infty} \frac{\sin(2\pi kft)}{k}$$

Thus, this waveform has an infinite number of frequency components and hence an infinite bandwidth. However, the peak amplitude of the *k*th frequency component, *kf*, is only $1/k$, so most of the energy in this waveform is in the first few frequency components. What happens if we limit the bandwidth to just the first three frequency components? We have already seen the answer, in Figure 3.7a. As we can



(a) $(4/\pi)\ [\sin(2\pi ft) + (1/3)\sin(2\pi(3f)t) + (1/5)\sin(2\pi(5f)t)]$

(b) $(4/\pi)\ [\sin(2\pi ft) + (1/3)\sin(2\pi(3f)t) + (1/5)\sin(2\pi(5f)t) + (1/7)\sin(2\pi(7f)t)]$

(c) $(4/\pi) \sum (1/k)\sin(2\pi(kf)t)$, for k odd

Figure 3.7 Frequency Components of Square Wave $(T = 1/f)$

see, the shape of the resulting waveform is reasonably close to that of the original square wave.

We can use Figures 3.4 and 3.7 to illustrate the relationship between data rate and bandwidth. Suppose that we are using a digital transmission system that is capable of transmitting signals with a bandwidth of 4 MHz. Let us attempt to transmit a sequence of alternating 1s and 0s as the square wave of Figure 3.7c. What data rate can be achieved? We look at three cases.

Case I. Let us approximate our square wave with the waveform of Figure 3.7a. Although this waveform is a "distorted" square wave, it is sufficiently close to the square wave that a receiver should be able to discriminate between a binary 0 and a binary 1. If we let $f = 10^6$ cycles/second $= 1$ MHz, then the bandwidth of the signal

$$s(t) = \frac{4}{\pi} \times \left[\sin((2\pi \times 10^6)t) + \frac{1}{3}\sin((2\pi \times 3 \times 10^6)t) + \frac{1}{5}\sin((2\pi \times 5 \times 10^6)t)\right]$$

is $(5 \times 10^6) - 10^6 = 4$ MHz. Note that for $f = 1$ MHz, the period of the fundamental frequency is $T = 1/10^6 = 10^{-6} = 1\ \mu$s. If we treat this waveform as a bit string of 1s and 0s, one bit occurs every 0.5 μs, for a data rate of $2 \times 10^6 = 2$ Mbps. Thus, for a bandwidth of 4 MHz, a data rate of 2 Mbps is achieved.

Case II. Now suppose that we have a bandwidth of 8 MHz. Let us look again at Figure 3.7a, but now with $f = 2$ MHz. Using the same line of reasoning as before, the bandwidth of the signal is $(5 \times 2 \times 10^6) - (2 \times 10^6) = 8$ MHz. But in this case $T = 1/f = 0.5\ \mu$s. As a result, one bit occurs every 0.25 μs for a data rate of 4 Mbps. Thus, other things being equal, by doubling the bandwidth, we double the potential data rate.

Case III. Now suppose that the waveform of Figure 3.4c is considered adequate for approximating a square wave. That is, the difference between a positive and negative pulse in Figure 3.4c is sufficiently distinct that the waveform can be successfully used to represent a sequence of 1s and 0s. Assume as in Case II that $f = 2$ MHz and $T = 1/f = 0.5\ \mu$s, so that one bit occurs every 0.25 μs for a data rate of 4 Mbps. Using the waveform of Figure 3.4c, the bandwidth of the signal is $(3 \times 2 \times 10^6) - (2 \times 10^6) = 4$ MHz. Thus, a given bandwidth can support various data rates depending on the ability of the receiver to discern the difference between 0 and 1 in the presence of noise and other impairments.

To summarize,

- **Case I:** Bandwidth = 4 MHz; data rate = 2 Mbps
- **Case II:** Bandwidth = 8 MHz; data rate = 4 Mbps
- **Case III:** Bandwidth = 4 MHz; data rate = 4 Mbps

We can draw the following conclusions from the preceding discussion. In general, any digital waveform will have infinite bandwidth. If we attempt to transmit this waveform as a signal over any medium, the transmission system will limit the bandwidth that can be transmitted. Furthermore, for any given medium, the greater the bandwidth transmitted, the greater the cost. Thus, on the one hand, economic and practical reasons dictate that digital information be approximated by a signal of limited bandwidth. On the other hand, limiting the bandwidth creates distortions, which makes the task of interpreting the received signal more difficult. The more limited the bandwidth, the greater the distortion, and the greater the potential for error by the receiver.

One more illustration should serve to reinforce these concepts. Figure 3.8 shows a digital bit stream with a data rate of 2000 bits per second. With a bandwidth of 2500 Hz, or even 1700 Hz, the representation is quite good. Furthermore, we can generalize these results. If the data rate of the digital signal is W bps, then a very

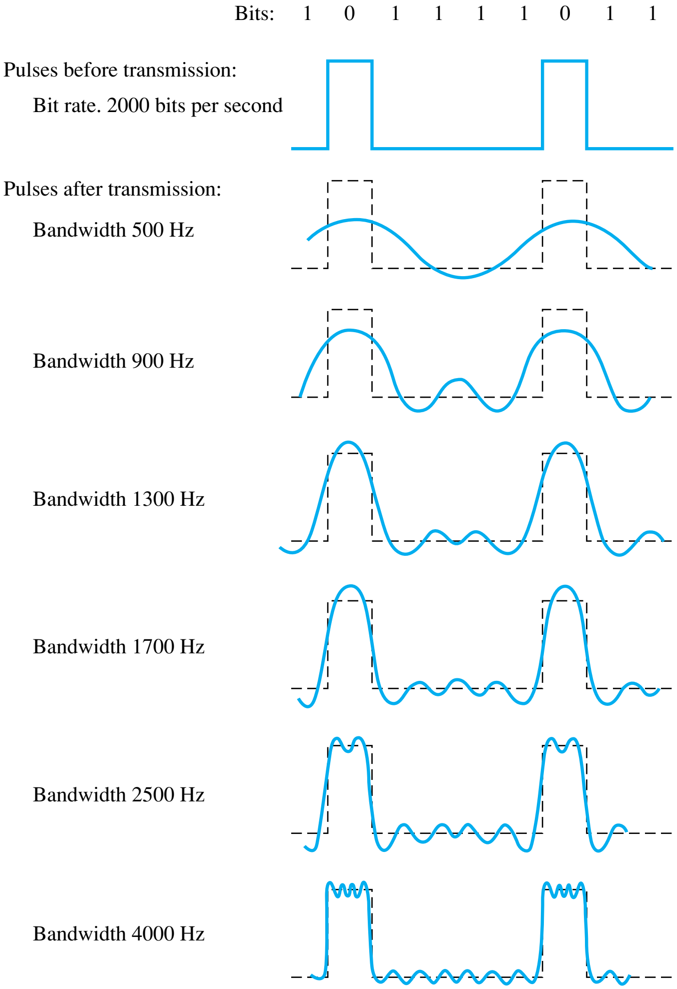


Figure 3.8 Effect of Bandwidth on a Digital Signal

good representation can be achieved with a bandwidth of $2W$ Hz. However, unless noise is very severe, the bit pattern can be recovered with less bandwidth than this (see the discussion of channel capacity in Section 3.4).

Thus, there is a direct relationship between data rate and bandwidth: The higher the data rate of a signal, the greater is its required effective bandwidth. Looked at the other way, the greater the bandwidth of a transmission system, the higher is the data rate that can be transmitted over that system.

Another observation worth making is this: If we think of the bandwidth of a signal as being centered about some frequency, referred to as the **center frequency**, then the higher the center frequency, the higher the potential bandwidth and therefore the higher the potential data rate. For example, if a signal is centered at 2 MHz, its maximum potential bandwidth is 4 MHz.

We return to a discussion of the relationship between bandwidth and data rate in Section 3.4, after a consideration of transmission impairments.

3.2 ANALOG AND DIGITAL DATA TRANSMISSION

The terms *analog* and *digital* correspond, roughly, to *continuous* and *discrete,* respectively. These two terms are used frequently in data communications in at least three contexts: data, signaling, and transmission.

Briefly, we define **data** as entities that convey meaning, or information. **Signals** are electric or electromagnetic representations of data. **Signaling** is the physical propagation of the signal along a suitable medium. **Transmission** is the communication of data by the propagation and processing of signals. In what follows, we try to make these abstract concepts clear by discussing the terms *analog* and *digital* as applied to data, signals, and transmission.

Analog and Digital Data

The concepts of analog and digital data are simple enough. Analog data take on continuous values in some interval. For example, voice and video are continuously varying patterns of intensity. Most data collected by sensors, such as temperature and pressure, are continuous valued. Digital data take on discrete values; examples are text and integers.

The most familiar example of analog data is **audio**, which, in the form of acoustic sound waves, can be perceived directly by human beings. Figure 3.9 shows the acoustic spectrum for human speech and for music.[5] Frequency components of typical speech may be found between approximately 100 Hz and 7 kHz. Although much of the energy in speech is concentrated at the lower frequencies, tests have shown that frequencies below 600 or 700 Hz add very little to the intelligibility of speech to the human ear. Typical speech has a dynamic range of about 25 dB;[6] that is, the power produced by the loudest shout may be as much as 300 times greater than the least whisper. Figure 3.9 also shows the acoustic spectrum and dynamic range for music.

Another common example of analog data is **video**. Here it is easier to characterize the data in terms of the TV screen (destination) rather than the original scene (source) recorded by the TV camera. To produce a picture on the screen, an electron beam scans across the surface of the screen from left to right and top to bottom. For black-and-white television, the amount of illumination produced (on a scale from black to white) at any point is proportional to the intensity of the beam as it passes that point. Thus at any instant in time the beam takes on an analog value of intensity to produce the desired brightness at that point on the screen. Further, as

[5] Note the use of a log scale for the *x*-axis. Because the *y*-axis is in units of decibels, it is effectively a log scale also. A basic review of log scales is in the math refresher document at the Computer Science Student Resource Site at WilliamStallings.com/StudentSupport.html.

[6] The concept of decibels is explained in Appendix 3A.

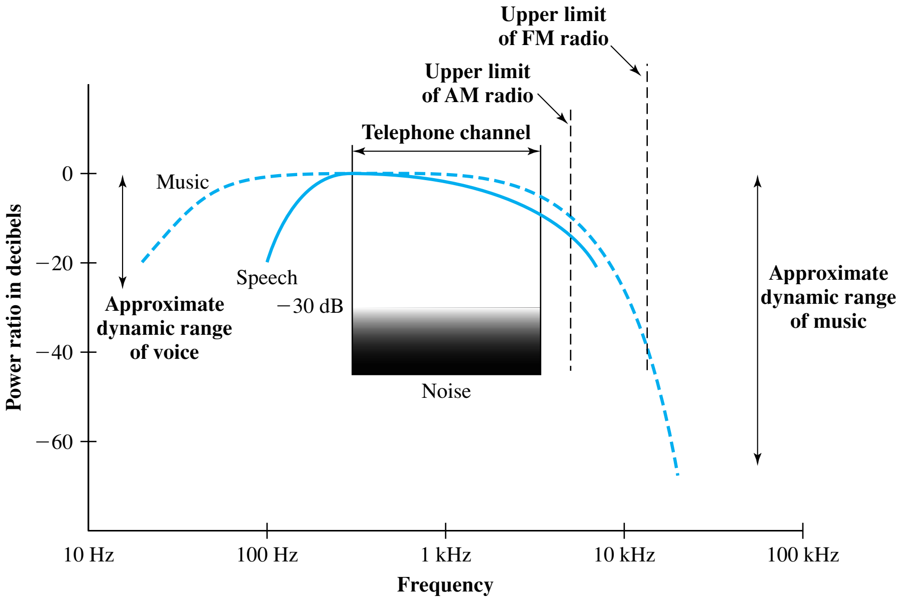


Figure 3.9 Acoustic Spectrum of Speech and Music [CARN99a]

the beam scans, the analog value changes. Thus the video image can be thought of as a time-varying analog signal.

Figure 3.10 depicts the scanning process. At the end of each scan line, the beam is swept rapidly back to the left (horizontal retrace). When the beam reaches the bottom, it is swept rapidly back to the top (vertical retrace). The beam is turned off (blanked out) during the retrace intervals.

To achieve adequate resolution, the beam produces a total of 483 horizontal lines at a rate of 30 complete scans of the screen per second. Tests have shown that this rate will produce a sensation of flicker rather than smooth motion. To provide a flicker-free image without increasing the bandwidth requirement, a technique known as **interlacing** is used. As Figure 3.10 shows, the odd numbered scan lines and the even numbered scan lines are scanned separately, with odd and even fields alternating on successive scans. The odd field is the scan from A to B and the even field is the scan from C to D. The beam reaches the middle of the screen's lowest line after 241.5 lines. At this point, the beam is quickly repositioned at the top of the screen and recommences in the middle of the screen's topmost visible line to produce an additional 241.5 lines interlaced with the original set. Thus the screen is refreshed 60 times per second rather than 30, and flicker is avoided.

A familiar example of digital data is **text** or character strings. While textual data are most convenient for human beings, they cannot, in character form, be easily stored or transmitted by data processing and communications systems. Such systems are designed for binary data. Thus a number of codes have been devised by which characters are represented by a sequence of bits. Perhaps the earliest common example of this is the Morse code. Today, the most commonly used text code is the

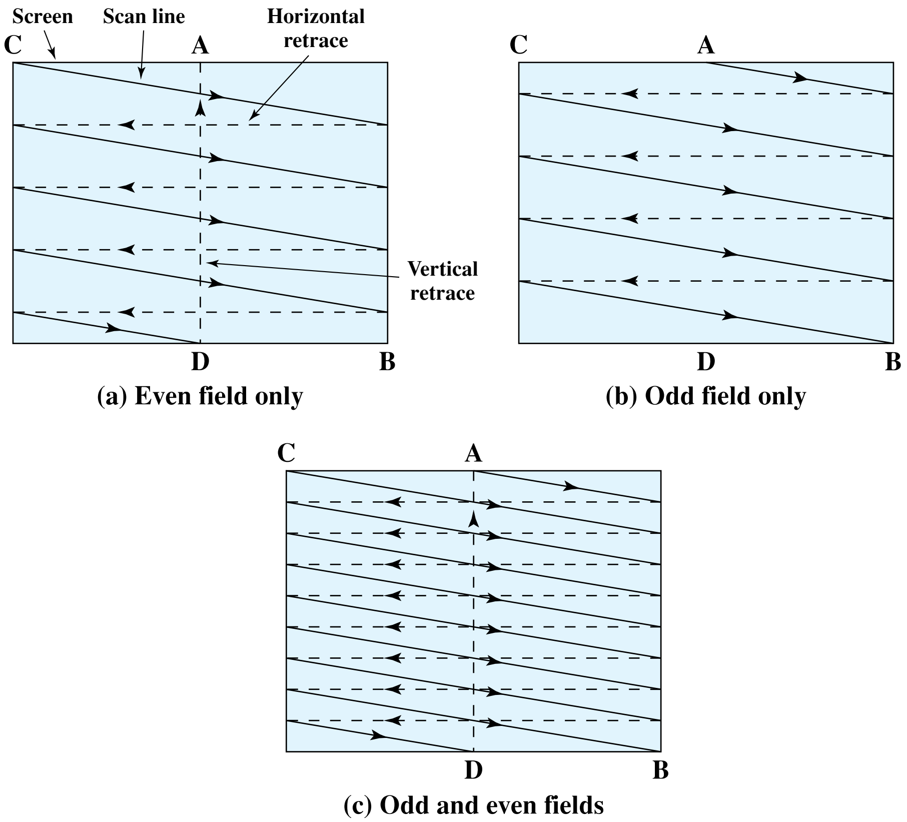


Figure 3.10 Video Interlaced Scanning

International Reference Alphabet (IRA).[7] Each character in this code is represented by a unique 7-bit pattern; thus 128 different characters can be represented. This is a larger number than is necessary, and some of the patterns represent invisible *control characters*. IRA-encoded characters are almost always stored and transmitted using 8 bits per character. The eighth bit is a parity bit used for error detection. This bit is set such that the total number of binary 1s in each octet is always odd (odd parity) or always even (even parity). Thus a transmission error that changes a single bit, or any odd number of bits, can be detected.

Analog and Digital Signals

In a communications system, data are propagated from one point to another by means of electromagnetic signals. An **analog signal** is a continuously varying electromagnetic wave that may be propagated over a variety of media, depending on

[7] IRA is defined in ITU-T Recommendation T.50 and was formerly known as International Alphabet Number 5 (IA5). The U.S. national version of IRA is referred to as the American Standard Code for Information Interchange (ASCII). Appendix E provides a description and table of the IRA code.

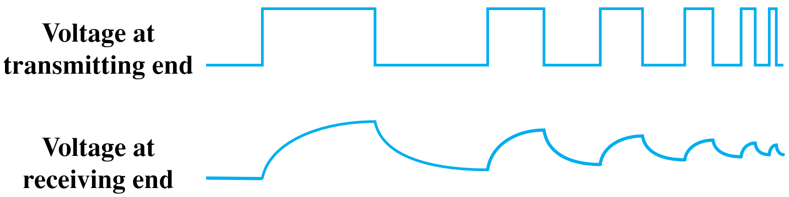


Figure 3.11 Attenuation of Digital Signals

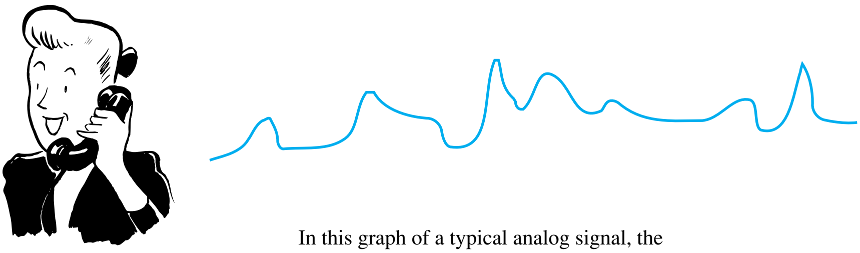
spectrum; examples are wire media, such as twisted pair and coaxial cable; fiber optic cable; and unguided media, such as atmosphere or space propagation. A **digital signal** is a sequence of voltage pulses that may be transmitted over a wire medium; for example, a constant positive voltage level may represent binary 0 and a constant negative voltage level may represent binary 1.

The principal advantages of digital signaling are that it is generally cheaper than analog signaling and is less susceptible to noise interference. The principal disadvantage is that digital signals suffer more from attenuation than do analog signals. Figure 3.11 shows a sequence of voltage pulses, generated by a source using two voltage levels, and the received voltage some distance down a conducting medium. Because of the attenuation, or reduction, of signal strength at higher frequencies, the pulses become rounded and smaller. It should be clear that this attenuation can lead rather quickly to the loss of the information contained in the propagated signal.

In what follows, we first look at some specific examples of signal types and then discuss the relationship between data and signals.

Examples Let us return to our three examples of the preceding subsection. For each example, we will describe the signal and estimate its bandwidth.

The most familiar example of analog information is **audio**, or acoustic, information, which, in the form of sound waves, can be perceived directly by human beings. One form of acoustic information, of course, is human speech. This form of information is easily converted to an electromagnetic signal for transmission (Figure 3.12). In essence, all of the sound frequencies, whose amplitude is measured in terms of loudness, are converted into electromagnetic frequencies, whose amplitude is measured in volts. The telephone handset contains a simple mechanism for making such a conversion.



In this graph of a typical analog signal, the variations in amplitude and frequency convey the gradations of loudness and pitch in speech or music. Similar signals are used to transmit television pictures, but at much higher frequencies.

Figure 3.12 Conversion of Voice Input to Analog Signal

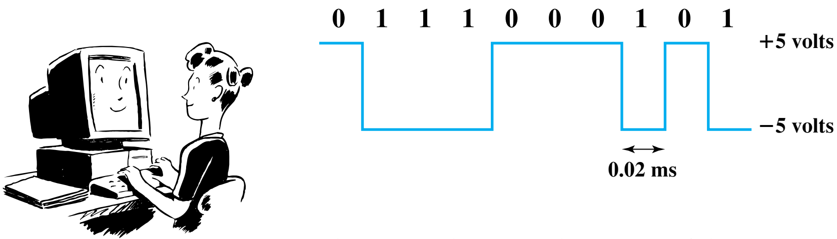
In the case of acoustic data (voice), the data can be represented directly by an electromagnetic signal occupying the same spectrum. However, there is a need to compromise between the fidelity of the sound as transmitted electrically and the cost of transmission, which increases with increasing bandwidth. As mentioned, the spectrum of speech is approximately 100 Hz to 7 kHz, although a much narrower bandwidth will produce acceptable voice reproduction. The standard spectrum for a voice channel is 300 to 3400 Hz. This is adequate for speech transmission, minimizes required transmission capacity, and allows the use of rather inexpensive telephone sets. The telephone transmitter converts the incoming acoustic voice signal into an electromagnetic signal over the range 300 to 3400 Hz. This signal is then transmitted through the telephone system to a receiver, which reproduces it as acoustic sound.

Now let us look at the **video** signal. To produce a video signal, a TV camera, which performs similar functions to the TV receiver, is used. One component of the camera is a photosensitive plate, upon which a scene is optically focused. An electron beam sweeps across the plate from left to right and top to bottom, in the same fashion as depicted in Figure 3.10 for the receiver. As the beam sweeps, an analog electric signal is developed proportional to the brightness of the scene at a particular spot. We mentioned that a total of 483 lines are scanned at a rate of 30 complete scans per second. This is an approximate number taking into account the time lost during the vertical retrace interval. The actual U.S. standard is 525 lines, but of these about 42 are lost during vertical retrace. Thus the horizontal scanning frequency is $(525 \text{ lines}) \times (30 \text{ scan/s}) = 15{,}750$ lines per second, or $63.5\ \mu\text{s/line}$. Of the $63.5\ \mu\text{s}$, about $11\ \mu\text{s}$ are allowed for horizontal retrace, leaving a total of $52.5\ \mu\text{s}$ per video line.

Now we are in a position to estimate the bandwidth required for the video signal. To do this we must estimate the upper (maximum) and lower (minimum) frequency of the band. We use the following reasoning to arrive at the maximum frequency: The maximum frequency would occur during the horizontal scan if the scene were alternating between black and white as rapidly as possible. We can estimate this maximum value by considering the resolution of the video image. In the vertical dimension, there are 483 lines, so the maximum vertical resolution would be 483. Experiments have shown that the actual subjective resolution is about 70% of that number, or about 338 lines. In the interest of a balanced picture, the horizontal and vertical resolutions should be about the same. Because the ratio of width to height of a TV screen is 4 : 3, the horizontal resolution should be about $4/3 \times 338 = 450$ lines. As a worst case, a scanning line would be made up of 450 elements alternating black and white. The scan would result in a wave, with each cycle of the wave consisting of one higher (black) and one lower (white) voltage level. Thus there would be $450/2 = 225$ cycles of the wave in $52.5\ \mu\text{s}$, for a maximum frequency of about 4.2 MHz. This rough reasoning, in fact, is fairly accurate. The lower limit is a dc or zero frequency, where the dc component corresponds to the average illumination of the scene (the average value by which the brightness exceeds the reference black level). Thus the bandwidth of the video signal is approximately $4 \text{ MHz} - 0 = 4 \text{ MHz}$.

The foregoing discussion did not consider color or audio components of the signal. It turns out that, with these included, the bandwidth remains about 4 MHz.

Finally, the third example described is the general case of **binary data**. Binary data is generated by terminals, computers, and other data processing equipment



User input at a PC is converted into a stream of binary digits (1s and 0s). In this graph of a typical digital signal, binary one is represented by −5 volts and binary zero is represented by +5 volts. The signal for each bit has a duration of 0.02 ms, giving a data rate of 50,000 bits per second (50 kbps).

Figure 3.13 Conversion of PC Input to Digital Signal

and then converted into digital voltage pulses for transmission, as illustrated in Figure 3.13. A commonly used signal for such data uses two constant (dc) voltage levels, one level for binary 1 and one level for binary 0. (In Chapter 5, we shall see that this is but one alternative, referred to as NRZ.) Again, we are interested in the bandwidth of such a signal. This will depend, in any specific case, on the exact shape of the waveform and the sequence of 1s and 0s. We can obtain some understanding by considering Figure 3.8 (compare Figure 3.7). As can be seen, the greater the bandwidth of the signal, the more faithfully it approximates a digital pulse stream.

Data and Signals In the foregoing discussion, we have looked at analog signals used to represent analog data and digital signals used to represent digital data. Generally, analog data are a function of time and occupy a limited frequency spectrum; such data can be represented by an electromagnetic signal occupying the same spectrum. Digital data can be represented by digital signals, with a different voltage level for each of the two binary digits.

As Figure 3.14 illustrates, these are not the only possibilities. Digital data can also be represented by analog signals by use of a modem (modulator/demodulator). The modem converts a series of binary (two-valued) voltage pulses into an analog signal by encoding the digital data onto a carrier frequency. The resulting signal occupies a certain spectrum of frequency centered about the carrier and may be propagated across a medium suitable for that carrier. The most common modems represent digital data in the voice spectrum and hence allow those data to be propagated over ordinary voice-grade telephone lines. At the other end of the line, another modem demodulates the signal to recover the original data.

In an operation very similar to that performed by a modem, analog data can be represented by digital signals. The device that performs this function for voice data is a codec (coder-decoder). In essence, the codec takes an analog signal that directly represents the voice data and approximates that signal by a bit stream. At the receiving end, the bit stream is used to reconstruct the analog data.

Thus, Figure 3.14 suggests that data may be encoded into signals in a variety of ways. We will return to this topic in Chapter 5.

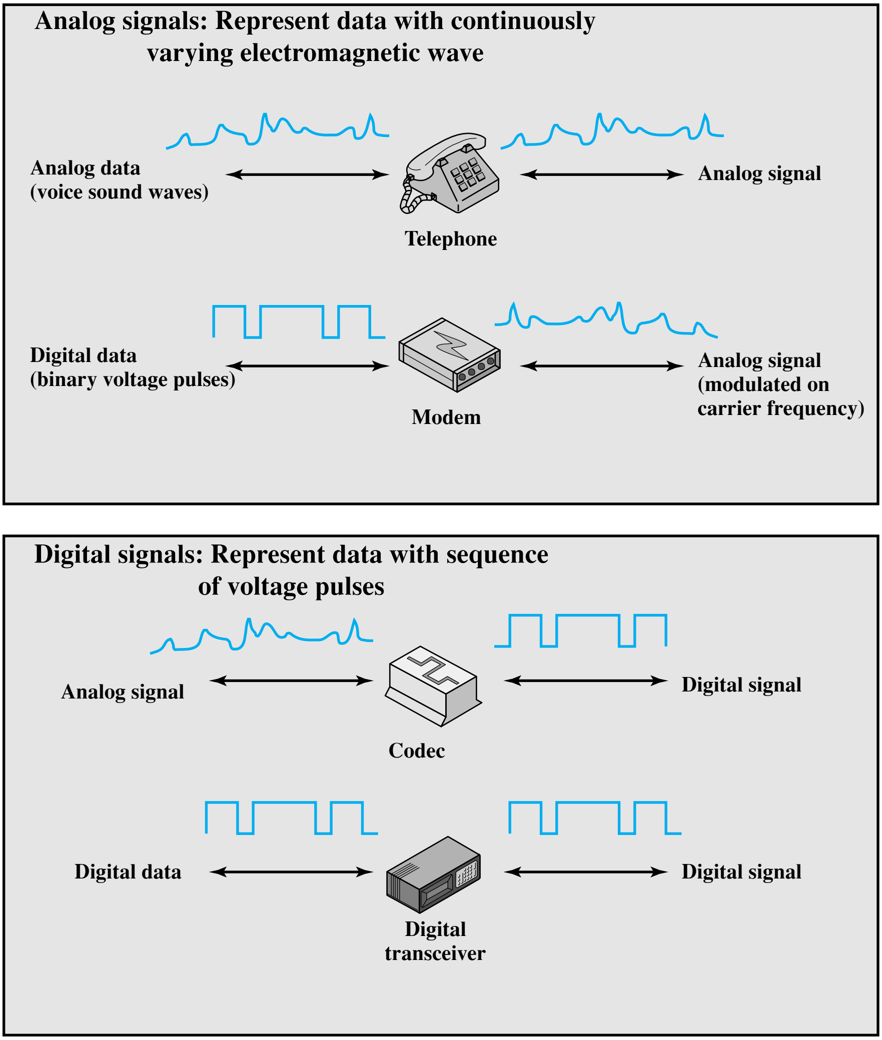

Analog signals: Represent data with continuously varying electromagnetic wave
Analog data (voice sound waves)
Telephone
Analog signal
Digital data (binary voltage pulses)
Modem
Analog signal (modulated on carrier frequency)
Digital signals: Represent data with sequence of voltage pulses
Analog signal
Codec
Digital signal
Digital data
Digital transceiver
Digital signal

Figure 3.14 Analog and Digital Signaling of Analog and Digital Data

Analog and Digital Transmission

Both analog and digital signals may be transmitted on suitable transmission media. The way these signals are treated is a function of the transmission system. Table 3.1 summarizes the methods of data transmission. **Analog transmission** is a means of transmitting analog signals without regard to their content; the signals may represent analog data (e.g., voice) or digital data (e.g., binary data that pass through a modem). In either case, the analog signal will become weaker (attenuate) after a certain distance. To achieve longer distances, the analog transmission system includes amplifiers that boost the energy in the signal. Unfortunately, the amplifier also boosts the noise components. With amplifiers cascaded to achieve long distances, the signal becomes more and more distorted.

Table 3.1 Analog and Digital Transmission

(a) Data and Signals

	Analog Signal	Digital Signal
Analog Data	Two alternatives: (1) signal occupies the same spectrum as the analog data; (2) analog data are encoded to occupy a different portion of spectrum.	Analog data are encoded using a codec to produce a digital bit stream.
Digital Data	Digital data are encoded using a modem to produce analog signal.	Two alternatives: (1) signal consists of two voltage levels to represent the two binary values; (2) digital data are encoded to produce a digital signal with desir ed properties.

(b) Treatment of Signals

	Analog Transmission	Digital Transmission
Analog Signal	Is propagated through amplifiers; same treatment whether signal is used to represent analog data or digital data.	Assumes that the analog signal represents digital data. Signal is propagated through repeaters; at each repeater, digital data are recovered from inbound signal and used to generate a new analog outbound signal.
Digital Signal	Not used	Digital signal represents a stream of 1s and 0s, which may represent digital data or may be an encoding of analog data. Signal is propagated through repeaters; at each repeater, stream of 1s and 0s is recovered from inbound signal and used to generate a new digital outbound signal.

For analog data, such as voice, quite a bit of distortion can be tolerated and the data remain intelligible. However, for digital data, cascaded amplifiers will introduce errors.

Digital transmission, in contrast, assumes a binary content to the signal. A digital signal can be transmitted only a limited distance before attenuation, noise, and other impairments endanger the integrity of the data. To achieve greater distances, repeaters are used. A repeater receives the digital signal, recovers the pattern of 1s and 0s, and retransmits a new signal. Thus the attenuation is overcome.

The same technique may be used with an analog signal if it is assumed that the signal carries digital data. At appropriately spaced points, the transmission system has repeaters rather than amplifiers. The repeater recovers the digital data from the analog signal and generates a new, clean analog signal. Thus noise is not cumulative.

The question naturally arises as to which is the preferred method of transmission. The answer being supplied by the telecommunications industry and its customers is digital. Both long-haul telecommunications facilities and intrabuilding services have moved to digital transmission and, where possible, digital signaling techniques. The most important reasons are as follows:

- **Digital technology:** The advent of large-scale integration (LSI) and very-large-scale integration (VLSI) technology has caused a continuing drop in the cost and size of digital circuitry. Analog equipment has not shown a similar drop.
- **Data integrity:** With the use of repeaters rather than amplifiers, the effects of noise and other signal impairments are not cumulative. Thus it is possible to transmit data longer distances and over lower quality lines by digital means while maintaining the integrity of the data.
- **Capacity utilization:** It has become economical to build transmission links of very high bandwidth, including satellite channels and optical fiber. A high degree of multiplexing is needed to utilize such capacity effectively, and this is more easily and cheaply achieved with digital (time division) rather than analog (frequency division) techniques. This is explored in Chapter 8.
- **Security and privacy:** Encryption techniques can be readily applied to digital data and to analog data that have been digitized.
- **Integration:** By treating both analog and digital data digitally, all signals have the same form and can be treated similarly. Thus economies of scale and convenience can be achieved by integrating voice, video, and digital data.

3.3 TRANSMISSION IMPAIRMENTS

With any communications system, the signal that is received may differ from the signal that is transmitted due to various transmission impairments. For analog signals, these impairments can degrade the signal quality. For digital signals, bit errors may be introduced, such that a binary 1 is transformed into a binary 0 or vice versa. In this section, we examine the various impairments and how they may affect the information-carrying capacity of a communication link; Chapter 5 looks at measures that can be taken to compensate for these impairments.

The most significant impairments are

- Attenuation and attenuation distortion
- Delay distortion
- Noise

Attenuation

The strength of a signal falls off with distance over any transmission medium. For guided media, this reduction in strength, or attenuation, is generally exponential and thus is typically expressed as a constant number of decibels per unit distance. For unguided media, attenuation is a more complex function of distance and the makeup of the atmosphere. Attenuation introduces three considerations for the transmission engineer. First, a received signal must have sufficient strength so that the electronic circuitry in the receiver can detect the signal. Second, the signal must maintain a level sufficiently higher than noise to be received without error. Third, attenuation varies with frequency.

The first and second problems are dealt with by attention to signal strength and the use of amplifiers or repeaters. For a point-to-point link, the signal strength of the transmitter must be strong enough to be received intelligibly, but not so strong as to overload the circuitry of the transmitter or receiver, which would cause distortion. Beyond a certain distance, the attenuation becomes unacceptably great, and repeaters or amplifiers are used to boost the signal at regular intervals. These problems are more complex for multipoint lines where the distance from transmitter to receiver is variable.

The third problem is particularly noticeable for analog signals. Because the attenuation varies as a function of frequency, the received signal is distorted, reducing intelligibility. To overcome this problem, techniques are available for equalizing attenuation across a band of frequencies. This is commonly done for voice-grade telephone lines by using loading coils that change the electrical properties of the line; the result is to smooth out attenuation effects. Another approach is to use amplifiers that amplify high frequencies more than lower frequencies.

An example is provided in Figure 3.15a, which shows attenuation as a function of frequency for a typical leased line. In the figure, attenuation is measured relative to the attenuation at 1000 Hz. Positive values on the *y*-axis represent attenuation greater than that at 1000 Hz. A 1000-Hz tone of a given power level is applied to the input, and the power, P_{1000}, is measured at the output. For any other frequency *f*, the procedure is repeated and the relative attenuation in decibels is[8]

$$N_f = -10 \log_{10} \frac{P_f}{P_{1000}}$$

The solid line in Figure 3.15a shows attenuation without equalization. As can be seen, frequency components at the upper end of the voice band are attenuated much more than those at lower frequencies. It should be clear that this will result in a distortion of the received speech signal. The dashed line shows the effect of equalization. The flattened response curve improves the quality of voice signals. It also allows higher data rates to be used for digital data that are passed through a modem.

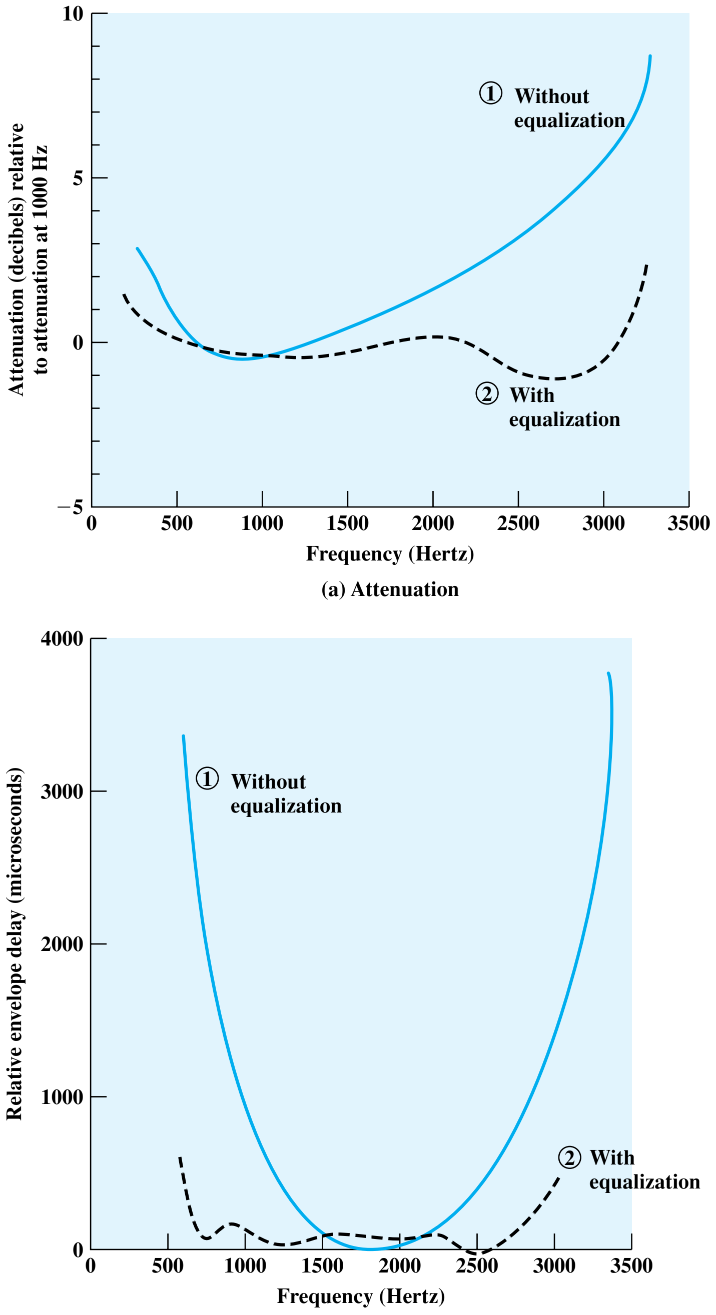
Attenuation distortion can present less of a problem with digital signals. As we have seen, the strength of a digital signal falls off rapidly with frequency (Figure 3.5b); most of the content is concentrated near the fundamental frequency or bit rate of the signal.

Delay Distortion

Delay distortion occurs because the velocity of propagation of a signal through a guided medium varies with frequency. For a bandlimited signal, the velocity tends to be highest near the center frequency and fall off toward the two edges of the band. Thus various frequency components of a signal will arrive at the receiver at different times, resulting in phase shifts between the different frequencies.

This effect is referred to as delay distortion because the received signal is distorted due to varying delays experienced at its constituent frequencies. Delay distortion is particularly critical for digital data. Consider that a sequence of bits is being transmitted, using either analog or digital signals. Because of delay distortion, some of the signal components of one bit position will spill over into other bit positions, causing **intersymbol interference**, which is a major limitation to maximum bit rate over a transmission channel.

[8] In the remainder of this book, unless otherwise indicated, we use log(x) to mean $\log_{10}(x)$.



(a) Attenuation

(b) Delay distortion

Figure 3.15 Attenuation and Delay Distortion Curves for a Voice Channel

Equalizing techniques can also be used for delay distortion. Again using a leased telephone line as an example, Figure 3.15b shows the effect of equalization on delay as a function of frequency.

Noise

For any data transmission event, the received signal will consist of the transmitted signal, modified by the various distortions imposed by the transmission system, plus additional unwanted signals that are inserted somewhere between transmission and reception. The latter, undesired signals are referred to as noise. Noise is the major limiting factor in communications system performance.

Noise may be divided into four categories:

- Thermal noise
- Intermodulation noise
- Crosstalk
- Impulse noise

Thermal noise is due to thermal agitation of electrons. It is present in all electronic devices and transmission media and is a function of temperature. Thermal noise is uniformly distributed across the bandwidths typically used in communications systems and hence is often referred to as **white noise**. Thermal noise cannot be eliminated and therefore places an upper bound on communications system performance. Because of the weakness of the signal received by satellite earth stations, thermal noise is particularly significant for satellite communication.

The amount of thermal noise to be found in a bandwidth of 1 Hz in any device or conductor is

$$N_0 = \mathrm{k}T(\mathrm{W/Hz})$$

where[9]

N_0 = noise power density in watts per 1 Hz of bandwidth

k = Boltzmann's constant = 1.38×10^{-23} J/K

T = temperature, in kelvins (absolute temperature), where the symbol K is used to represent 1 kelvin

EXAMPLE 3.1 Room temperature is usually specified as $T = 17°\mathrm{C}$, or 290 K. At this temperature, the thermal noise power density is

$$N_0 = (1.38 \times 10^{-23}) \times 290 = 4 \times 10^{-21}\ \mathrm{W/Hz} = -204\ \mathrm{dBW/Hz}$$

where dBW is the decibel-watt, defined in Appendix 3A.

[9] A Joule (J) is the International System (SI) unit of electrical, mechanical, and thermal energy. A Watt is the SI unit of power, equal to one Joule per second. The kelvin (K) is the SI unit of thermodynamic temperature. For a temperature in kelvins of T, the corresponding temperature in degrees Celsius is equal to $T - 273.15$.

The noise is assumed to be independent of frequency. Thus the thermal noise in watts present in a bandwidth of B Hertz can be expressed as

$$N = \mathrm{k}TB$$

or, in decibel-watts,

$$\begin{aligned} N &= 10 \log k + 10 \log T + 10 \log B \\ &= -228.6 \text{ dBW} + 10 \log T + 10 \log B \end{aligned}$$

EXAMPLE 3.2 Given a receiver with an effective noise temperature of 294 K and a 10-MHz bandwidth, the thermal noise level at the receiver's output is

$$\begin{aligned} N &= -228.6 \text{ dBW} + 10 \log(294) + 10 \log 10^7 \\ &= -228.6 + 24.7 + 70 \\ &= -133.9 \text{ dBW} \end{aligned}$$

When signals at different frequencies share the same transmission medium, the result may be **intermodulation noise**. The effect of intermodulation noise is to produce signals at a frequency that is the sum or difference of the two original frequencies or multiples of those frequencies. For example, the mixing of signals at frequencies f_1 and f_2 might produce energy at the frequency $f_1 + f_2$. This derived signal could interfere with an intended signal at the frequency $f_1 + f_2$.

Intermodulation noise is produced by nonlinearities in the transmitter, receiver, and/or intervening transmission medium. Ideally, these components behave as linear systems; that is, the output is equal to the input times a constant. However, in any real system, the output is a more complex function of the input. Excessive nonlinearity can be caused by component malfunction or overload from excessive signal strength. It is under these circumstances that the sum and difference frequency terms occur.

Crosstalk has been experienced by anyone who, while using the telephone, has been able to hear another conversation; it is an unwanted coupling between signal paths. It can occur by electrical coupling between nearby twisted pairs or, rarely, coax cable lines carrying multiple signals. Crosstalk can also occur when microwave antennas pick up unwanted signals; although highly directional antennas are used, microwave energy does spread during propagation. Typically, crosstalk is of the same order of magnitude as, or less than, thermal noise.

All of the types of noise discussed so far have reasonably predictable and relatively constant magnitudes. Thus it is possible to engineer a transmission system to cope with them. **Impulse noise**, however, is noncontinuous, consisting of irregular pulses or noise spikes of short duration and of relatively high amplitude. It is generated from a variety of causes, including external electromagnetic disturbances, such as lightning, and faults and flaws in the communications system.

Impulse noise is generally only a minor annoyance for analog data. For example, voice transmission may be corrupted by short clicks and crackles with no loss of intelligibility. However, impulse noise is the primary source of error in digital data

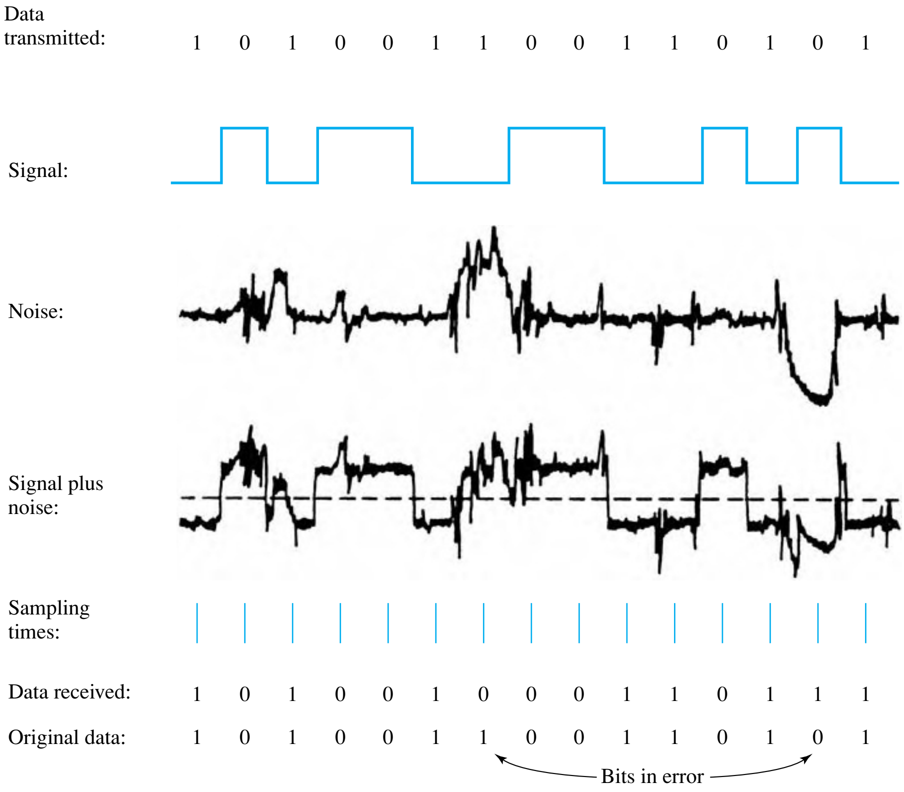


Figure 3.16 Effect of Noise on a Digital Signal

communication. For example, a sharp spike of energy of 0.01 s duration would not destroy any voice data but would wash out about 560 bits of digital data being transmitted at 56 kbps. Figure 3.16 is an example of the effect of noise on a digital signal. Here the noise consists of a relatively modest level of thermal noise plus occasional spikes of impulse noise. The digital data can be recovered from the signal by sampling the received waveform once per bit time. As can be seen, the noise is occasionally sufficient to change a 1 to a 0 or a 0 to a 1.

3.4 CHANNEL CAPACITY

We have seen that there are a variety of impairments that distort or corrupt a signal. For digital data, the question that then arises is to what extent these impairments limit the data rate that can be achieved. The maximum rate at which data can be transmitted over a given communication path, or channel, under given conditions, is referred to as the **channel capacity**.

There are four concepts here that we are trying to relate to one another.

- **Data rate:** The rate, in bits per second (bps), at which data can be communicated

- **Bandwidth:** The bandwidth of the transmitted signal as constrained by the transmitter and the nature of the transmission medium, expressed in cycles per second, or Hertz
- **Noise:** The average level of noise over the communications path
- **Error rate:** The rate at which errors occur, where an error is the reception of a 1 when a 0 was transmitted or the reception of a 0 when a 1 was transmitted

The problem we are addressing is this: Communications facilities are expensive and, in general, the greater the bandwidth of a facility, the greater the cost. Furthermore, all transmission channels of any practical interest are of limited bandwidth. The limitations arise from the physical properties of the transmission medium or from deliberate limitations at the transmitter on the bandwidth to prevent interference from other sources. Accordingly, we would like to make as efficient use as possible of a given bandwidth. For digital data, this means that we would like to get as high a data rate as possible at a particular limit of error rate for a given bandwidth. The main constraint on achieving this efficiency is noise.

Nyquist Bandwidth

To begin, let us consider the case of a channel that is noise free. In this environment, the limitation on data rate is simply the bandwidth of the signal. A formulation of this limitation, due to Nyquist, states that if the rate of signal transmission is $2B$, then a signal with frequencies no greater than B is sufficient to carry the signal rate. The converse is also true: Given a bandwidth of B, the highest signal rate that can be carried is $2B$. This limitation is due to the effect of intersymbol interference, such as is produced by delay distortion. The result is useful in the development of digital-to-analog encoding schemes and is, in essence, based on the same derivation as that of the sampling theorem, described in Appendix F.

Note that in the preceding paragraph, we referred to signal rate. If the signals to be transmitted are binary (two voltage levels), then the data rate that can be supported by B Hz is $2B$ bps. However, as we shall see in Chapter 5, signals with more than two levels can be used; that is, each signal element can represent more than one bit. For example, if four possible voltage levels are used as signals, then each signal element can represent two bits. With multilevel signaling, the Nyquist formulation becomes

$$C = 2B \log_2 M$$

where M is the number of discrete signal or voltage levels.

So, for a given bandwidth, the data rate can be increased by increasing the number of different signal elements. However, this places an increased burden on the receiver: Instead of distinguishing one of two possible signal elements during each signal time, it must distinguish one of M possible signal elements. Noise and other impairments on the transmission line will limit the practical value of M.

EXAMPLE 3.3 Consider a voice channel being used, via modem, to transmit digital data. Assume a bandwidth of 3100 Hz. Then the Nyquist capacity, C, of the channel is $2B = 6200$ bps. For $M = 8$, a value used with some modems, C becomes 18,600 bps for a bandwidth of 3100 Hz.

Shannon Capacity Formula

Nyquist's formula indicates that, all other things being equal, doubling the bandwidth doubles the data rate. Now consider the relationship among data rate, noise, and error rate. The presence of noise can corrupt one or more bits. If the data rate is increased, then the bits become "shorter" so that more bits are affected by a given pattern of noise.

Figure 3.16 illustrates this relationship. If the data rate is increased, then more bits will occur during the interval of a noise spike, and hence more errors will occur.

All of these concepts can be tied together neatly in a formula developed by the mathematician Claude Shannon. As we have just illustrated, the higher the data rate, the more damage that unwanted noise can do. For a given level of noise, we would expect that a greater signal strength would improve the ability to receive data correctly in the presence of noise. The key parameter involved in this reasoning is the **signal-to-noise ratio** (SNR, or S/N),[10] which is the ratio of the power in a signal to the power contained in the noise that is present at a particular point in the transmission. Typically, this ratio is measured at a receiver, because it is at this point that an attempt is made to process the signal and recover the data. For convenience, this ratio is often reported in decibels:

$$\mathrm{SNR}_{\mathrm{dB}} = 10 \log_{10} \frac{\text{signal power}}{\text{noise power}}$$

This expresses the amount, in decibels, that the intended signal exceeds the noise level. A high SNR will mean a high-quality signal and a low number of required intermediate repeaters.

The signal-to-noise ratio is important in the transmission of digital data because it sets the upper bound on the achievable data rate. Shannon's result is that the maximum channel capacity, in bits per second, obeys the equation

$$C = B \log_2(1 + \mathrm{SNR}) \tag{3.1}$$

where C is the capacity of the channel in bits per second and B is the bandwidth of the channel in Hertz. The Shannon formula represents the theoretical maximum that can be achieved. In practice, however, only much lower rates are achieved. One reason for this is that the formula assumes white noise (thermal noise). Impulse noise is not accounted for, nor are attenuation distortion or delay distortion. Even in

[10] Some of the literature uses SNR; others use S/N. Also, in some cases the dimensionless quantity is referred to as SNR or S/N and the quantity in decibels is referred to as $\mathrm{SNR}_{\mathrm{db}}$ or $(\mathrm{S/N})_{\mathrm{db}}$. Others use just SNR or S/N to mean the dB quantity. This text uses SNR and $\mathrm{SNR}_{\mathrm{db}}$.

an ideal white noise environment, present technology still cannot achieve Shannon capacity due to encoding issues, such as coding length and complexity.

The capacity indicated in the preceding equation is referred to as the error-free capacity. Shannon proved that if the actual information rate on a channel is less than the error-free capacity, then it is theoretically possible to use a suitable signal code to achieve error-free transmission through the channel. Shannon's theorem unfortunately does not suggest a means for finding such codes, but it does provide a yardstick by which the performance of practical communication schemes may be measured.

Several other observations concerning the preceding equation may be instructive. For a given level of noise, it would appear that the data rate could be increased by increasing either signal strength or bandwidth. However, as the signal strength increases, so do the effects of nonlinearities in the system, leading to an increase in intermodulation noise. Note also that, because noise is assumed to be white, the wider the bandwidth, the more noise is admitted to the system. Thus, as B increases, SNR decreases.

EXAMPLE 3.4 Let us consider an example that relates the Nyquist and Shannon formulations. Suppose that the spectrum of a channel is between 3 MHz and 4 MHz and $\mathrm{SNR_{dB}} = 24$ dB. Then

$$B = 4\text{ MHz} - 3\text{ MHz} = 1\text{ MHz}$$

$$\mathrm{SNR_{dB}} = 24\text{ dB} = 10\log_{10}(\mathrm{SNR})$$

$$\mathrm{SNR} = 251$$

Using Shannon's formula,

$$C = 10^6 \times \log_2(1 + 251) \approx 10^6 \times 8 = 8\text{ Mbps}$$

This is a theoretical limit and, as we have said, is unlikely to be reached. But assume we can achieve the limit. Based on Nyquist's formula, how many signaling levels are required? We have

$$C = 2B\log_2 M$$

$$8 \times 10^6 = 2 \times (10^6) \times \log_2 M$$

$$4 = \log_2 M$$

$$M = 16$$

The Expression E_b/N_0

Finally, we mention a parameter related to SNR that is more convenient for determining digital data rates and error rates and that is the standard quality measure for digital communication system performance. The parameter is the ratio of signal energy per bit to noise power density per Hertz, E_b/N_0. Consider a signal, digital or analog, that contains binary digital data transmitted at a certain bit rate R. Recalling that 1 Watt = 1 J/s, the energy per bit in a signal is given by $E_b = ST_b$, where S is the signal power and T_b is the time required to send one bit. The data rate R is just $R = 1/T_b$. Thus

$$\frac{E_b}{N_0} = \frac{S/R}{N_0} = \frac{S}{kTR}$$

or, in decibel notation,

$$\left(\frac{E_b}{N_0}\right)_{\mathrm{dB}} = S_{\mathrm{dBW}} - 10 \log R - 10 \log \mathrm{k} - 10 \log T$$
$$= S_{\mathrm{dBW}} - 10 \log R + 228.6\ \mathrm{dBW} - 10 \log T$$

The ratio E_b/N_0 is important because the bit error rate for digital data is a (decreasing) function of this ratio. Given a value of E_b/N_0 needed to achieve a desired error rate, the parameters in the preceding formula may be selected. Note that as the bit rate R increases, the transmitted signal power, relative to noise, must increase to maintain the required E_b/N_0.

Let us try to grasp this result intuitively by considering again Figure 3.16. The signal here is digital, but the reasoning would be the same for an analog signal. In several instances, the noise is sufficient to alter the value of a bit. If the data rate were doubled, the bits would be more tightly packed together, and the same passage of noise might destroy two bits. Thus, for constant signal to noise ratio, an increase in data rate increases the error rate.

The advantage of E_b/N_0 over SNR is that the latter quantity depends on the bandwidth.

EXAMPLE 3.5 For binary phase-shift keying (defined in Chapter 5), $E_b/N_0 =$ 8.4 dB is required for a bit error rate of 10^{-4} (one bit error out of every 10,000). If the effective noise temperature is 290°K (room temperature) and the data rate is 2400 bps, what received signal level is required?

We have

$$\begin{aligned} 8.4 &= S(\mathrm{dBW}) - 10 \log 2400 + 228.6\ \mathrm{dBW} - 10 \log 290 \\ &= S(\mathrm{dBW}) - (10)(3.38) + 228.6 - (10)(2.46) \\ S &= -161.8\ \mathrm{dBW} \end{aligned}$$

We can relate E_b/N_0 to SNR as follows. We have

$$\frac{E_b}{N_0} = \frac{S}{N_0 R}$$

The parameter N_0 is the noise power density in Watts/Hertz. Hence, the noise in a signal with bandwidth B is $N = N_0 B$. Substituting, we have

$$\frac{E_b}{N_0} = \frac{S}{N}\frac{B_T}{R} \tag{3.2}$$

Another formulation of interest relates E_b/N_0 to spectral efficiency. Shannon's result (Equation 3.1) can be rewritten as:

$$\frac{S}{N} = 2^{C/B} - 1$$

Using Equation (3.2), and equating R with C, we have

$$\frac{E_b}{N_0} = \frac{B}{C}(2^{C/B} - 1)$$

This is a useful formula that relates the achievable spectral efficiency C/B to E_b/N_0.

EXAMPLE 3.6 Suppose we want to find the minimum E_b/N_0 required to achieve a spectral efficiency of 6 bps/Hz. Then

$$E_b/N_0 = (1/6)(2^6 - 1) = 10.5 = 10.21 \text{ dB}.$$

3.5 RECOMMENDED READING AND WEB SITE

There are many books that cover the fundamentals of analog and digital transmission. [COUC01] is quite thorough. Other good reference works are [FREE05], which includes some of the examples used in this chapter, and [HAYK01].

COUC01 Couch, L. *Digital and Analog Communication Systems*. Upper Saddle River, NJ: Prentice Hall, 2001.

FREE05 Freeman, R. *Fundamentals of Telecommunications*. New York: Wiley, 2005.

HAYK01 Haykin, S. *Communication Systems*. New York: Wiley, 2001.

Recommended Web site:

- **Fourier series synthesis:** An excellent visualization tool for Fourier series

3.6 KEY TERMS, REVIEW QUESTIONS, AND PROBLEMS

Key Terms

absolute bandwidth
analog data
analog signal
analog transmission
aperiodic
attenuation
attenuation distortion
audio
bandwidth
center frequency
channel capacity
crosstalk
data
dc component
decibel (dB)
delay distortion
digital data
digital signal

digital transmission
direct link
effective bandwidth
frequency
frequency domain
full duplex
fundamental frequency
gain
guided media
half duplex
impulse noise
interlacing
intermodulation noise
loss
multipoint link
noise
Nyquist bandwidth
peak amplitude
period
periodic signal
point-to-point link
phase
signal
signal-to-noise ratio (SNR)
signaling
simplex
sinusoid
spectrum
thermal noise
time domain
transmission
unguided media
video
wavelength
wireless

Review Questions

3.1. Differentiate between guided media and unguided media.
3.2. Differentiate between an analog and a digital electromagnetic signal.
3.3. What are three important characteristics of a periodic signal?
3.4. How many radians are there in a complete circle of 360 degrees?
3.5. What is the relationship between the wavelength and frequency of a sine wave?
3.6. Define *fundamental frequency*.
3.7. What is the relationship between a signal's spectrum and its bandwidth?
3.8. What is attenuation?
3.9. Define *channel capacity*.
3.10. What key factors affect channel capacity?

Problems

3.1 **a.** For multipoint configuration, only one device at a time can transmit. Why?
b. There are two methods of enforcing the rule that only one device can transmit. In the centralized method, one station is in control and can either transmit or allow a specified other station to transmit. In the decentralized method, the stations jointly cooperate in taking turns. What do you see as the advantages and disadvantages of the two methods?

3.2 A signal has a fundamental frequency of 1000 Hz. What is its period?

3.3 Express the following in the simplest form you can:
a. $\sin(2\pi ft - \pi) + \sin(2\pi ft + \pi)$
b. $\sin 2\pi ft + \sin(2\pi ft - \pi)$

3.4 Sound may be modeled as sinusoidal functions. Compare the relative frequency and wavelength of musical notes. Use 330 m/s as the speed of sound and the following frequencies for the musical scale.

Note	C	D	E	F	G	A	B	C
Frequency	264	297	330	352	396	440	495	528

3.5 If the solid curve in Figure 3.17 represents $\sin(2\pi t)$, what does the dotted curve represent? That is, the dotted curve can be written in the form $A \sin(2\pi ft + \phi)$; what are A, f, and ϕ?

3.6 Decompose the signal $(1 + 0.1 \cos 5t)\cos 100t$ into a linear combination of sinusoidal functions, and find the amplitude, frequency, and phase of each component. *Hint:* Use the identity for $\cos a \cos b$.

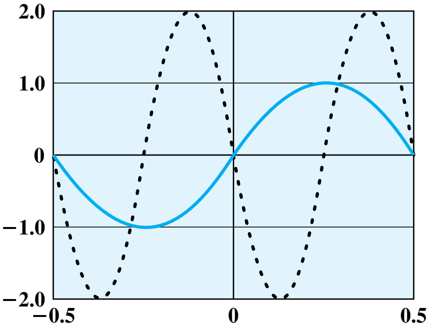


Figure 3.17 Figure for Problem 3.5

3.7 Find the period of the function $f(t) = (10 \cos t)^2$.

3.8 Consider two periodic functions $f_1(t)$ and $f_2(t)$, with periods T_1 and T_2, respectively. Is it always the case that the function $f(t) = f_1(t) + f_2(t)$ is periodic? If so, demonstrate this fact. If not, under what conditions is $f(t)$ periodic?

3.9 Figure 3.4 shows the effect of eliminating higher-harmonic components of a square wave and retaining only a few lower harmonic components. What would the signal look like in the opposite case; that is, retaining all higher harmonics and eliminating a few lower harmonics?

3.10 Figure 3.5b shows the frequency domain function for a single square pulse. The single pulse could represent a digital 1 in a communication system. Note that an infinite number of higher frequencies of decreasing magnitudes is needed to represent the single pulse. What implication does that have for a real digital transmission system?

3.11 IRA is a 7-bit code that allows 128 characters to be defined. In the 1970s, many newspapers received stories from the wire services in a 6-bit code called TTS. This code carried upper- and lower case characters as well as many special characters and formatting commands. The typical TTS character set allowed over 100 characters to be defined. How do you think this could be accomplished?

3.12 For a video signal, what increase in horizontal resolution is possible if a bandwidth of 5 MHz is used? What increase in vertical resolution is possible? Treat the two questions separately; that is, the increased bandwidth is to be used to increase either horizontal or vertical resolution, but not both.

3.13 **a.** Suppose that a digitized TV picture is to be transmitted from a source that uses a matrix of 480 × 500 picture elements (pixels), where each pixel can take on one of 32 intensity values. Assume that 30 pictures are sent per second. (This digital source is roughly equivalent to broadcast TV standards that have been adopted.) Find the source rate R (bps).

b. Assume that the TV picture is to be transmitted over a channel with 4.5-MHz bandwidth and a 35-dB signal-to-noise ratio. Find the capacity of the channel (bps).

c. Discuss how the parameters given in part (a) could be modified to allow transmission of color TV signals without increasing the required value for R.

3.14 Given an amplifier with an effective noise temperature of 10,000 K and a 10-MHz bandwidth, what thermal noise level, in dBW, may we expect at its output?

3.15 What is the channel capacity for a teleprinter channel with a 300-Hz bandwidth and a signal-to-noise ratio of 3 dB, where the noise is white thermal noise?

3.16 A digital signaling system is required to operate at 9600 bps.

a. If a signal element encodes a 4-bit word, what is the minimum required bandwidth of the channel?

b. Repeat part (a) for the case of 8-bit words.

3.17 What is the thermal noise level of a channel with a bandwidth of 10 kHz carrying 1000 watts of power operating at 50°C?

3.18 Given the narrow (usable) audio bandwidth of a telephone transmission facility, a nominal SNR of 56dB (400,000), and a certain level of distortion,

a. What is the theoretical maximum channel capacity (kbps) of traditional telephone lines?

b. What can we say about the actual maximum channel capacity?

3.19 Study the works of Shannon and Nyquist on channel capacity. Each places an upper limit on the bit rate of a channel based on two different approaches. How are the two related?

3.19 Consider a channel with a 1-MHz capacity and an SNR of 63.

a. What is the upper limit to the data rate that the channel can carry?

b. The result of part (a) is the upper limit. However, as a practical matter, better error performance will be achieved at a lower data rate. Assume we choose a data rate of 2/3 the maximum theoretical limit. How many signal levels are needed to achieve this data rate?

3.20 Given the narrow (usable) audio bandwidth of a telephone transmission facility, a nominal SNR_{dB} of 56dB (400,000), and a distortion level of <0.2%,

a. What is the theoretical maximum channel capacity (kbps) of traditional telephone lines?

b. What is the actual maximum channel capacity?

3.21 Given a channel with an intended capacity of 20 Mbps, the bandwidth of the channel is 3 MHz. Assuming white thermal noise, what signal-to-noise ratio is required to achieve this capacity?

3.22 The square wave of Figure 3.7c, with $T = 1$ ms, is passed through a lowpass filter that passes frequencies up to 8 kHz with no attenuation.

a. Find the power in the output waveform.

b. Assuming that at the filter input there is a thermal noise voltage with $N_0 = 0.1\ \mu$Watt/Hz, find the output signal to noise ratio in dB.

3.23 If the received signal level for a particular digital system is −151 dBW and the receiver system effective noise temperature is 1500 K, what is E_b/N_0 for a link transmitting 2400 bps?

3.24 Fill in the missing elements in the following table of approximate power ratios for various dB levels.

Decibels	1	2	3	4	5	6	7	8	9	10
Losses			0.5							0.1
Gains			2						10	

3.25 If an amplifier has a 30-dB voltage gain, what voltage ratio does the gain represent?

3.26 An amplifier has an output of 20 W. What is its output in dBW?

APPENDIX 3A DECIBELS AND SIGNAL STRENGTH

An important parameter in any transmission system is the signal strength. As a signal propagates along a transmission medium, there will be a loss, or *attenuation,* of signal strength. To compensate, amplifiers may be inserted at various points to impart a gain in signal strength.

It is customary to express gains, losses, and relative levels in decibels because

- Signal strength often falls off exponentially, so loss is easily expressed in terms of the decibel, which is a logarithmic unit.
- The net gain or loss in a cascaded transmission path can be calculated with simple addition and subtraction.

Table 3.2 Decibel Values

Power Ratio	dB	Power Ratio	dB
10^1	10	10^{-1}	−10
10^2	20	10^{-2}	−20
10^3	30	10^{-3}	−30
10^4	40	10^{-4}	−40
10^5	50	10^{-5}	−50
10^6	60	10^{-6}	−60

The decibel is a measure of the ratio between two signal levels. The decibel gain is given by

$$G_{dB} = 10 \log_{10} \frac{P_{out}}{P_{in}}$$

where

$$\begin{aligned} G_{dB} &= \text{gain, in decibels} \\ P_{in} &= \text{input power level} \\ P_{out} &= \text{output power level} \\ \log_{10} &= \text{logarithm to the base 10} \end{aligned}$$

Table 3.2 shows the relationship between decibel values and powers of 10.

There is some inconsistency in the literature over the use of the terms **gain** and **loss**. If the value of G_{dB} is positive, this represents an actual gain in power. For example, a gain of 3 dB means that the power has doubled. If the value of G_{dB} is negative, this represents an actual loss in power. For example, a gain of −3 dB means that the power has halved, and this is a loss of power. Normally, this is expressed by saying there is a loss of 3 dB. However, some of the literature would say that this is a loss of −3 dB. It makes more sense to say that a negative gain corresponds to a positive loss. Therefore, we define a decibel loss as

$$L_{dB} = -10 \log_{10} \frac{P_{out}}{P_{in}} = 10 \log_{10} \frac{P_{in}}{P_{out}} \qquad \textbf{(3.3)}$$

EXAMPLE 3.7 If a signal with a power level of 10 mW is inserted onto a transmission line and the measured power some distance away is 5 mW, the loss can be expressed as

$$L_{dB} = 10 \log(10/5) = 10(0.3) = 3 \text{ dB}.$$

Note that the decibel is a measure of relative, not absolute, difference. A loss from 1000 mW to 500 mW is also a loss of 3 dB.

The decibel is also used to measure the difference in voltage, taking into account that power is proportional to the square of the voltage:

$$P = \frac{V^2}{R}$$

where

$$\begin{aligned} P &= \text{power dissipated across resistance } R \\ V &= \text{voltage across resistance } R \end{aligned}$$

Thus

$$L_{\text{dB}} = 10 \log \frac{P_{\text{in}}}{P_{\text{out}}} = 10 \log \frac{V_{\text{in}}^2/R}{V_{\text{out}}^2/R} = 20 \log \frac{V_{\text{in}}}{V_{\text{out}}}$$

EXAMPLE 3.8 Decibels are useful in determining the gain or loss over a series of transmission elements. Consider a series in which the input is at a power level of 4 mW, the first element is a transmission line with a 12-dB loss (−12-dB gain), the second element is an amplifier with a 35-dB gain, and the third element is a transmission line with a 10-dB loss. The net gain is $(-12 + 35 - 10) = 13$ dB. To calculate the output power P_{out}:

$$\begin{aligned} G_{\text{dB}} &= 13 = 10 \log(P_{\text{out}}/4 \text{ mW}) \\ P_{\text{out}} &= 4 \times 10^{1.3} \text{ mW} = 79.8 \text{ mW} \end{aligned}$$

Decibel values refer to relative magnitudes or changes in magnitude, not to an absolute level. It is convenient to be able to refer to an absolute level of power or voltage in decibels so that gains and losses with reference to an initial signal level may be calculated easily. The **dBW (decibel-Watt)** is used extensively in microwave applications. The value of 1 W is selected as a reference and defined to be 0 dBW. The absolute decibel level of power in dBW is defined as

$$\text{Power}_{\text{dBW}} = 10 \log \frac{\text{Power}_{\text{W}}}{1 \text{ W}}$$

EXAMPLE 3.9 A power of 1000 W is 30 dBW, and a power of 1 mW is −30 dBW.

Another common unit is the **dBm (decibel-milliWatt)**, which uses 1 mW as the reference. Thus 0 dBm = 1 mW. The formula is

$$\text{Power}_{\text{dBm}} = 10 \log \frac{\text{Power}_{\text{mW}}}{1 \text{ mW}}$$

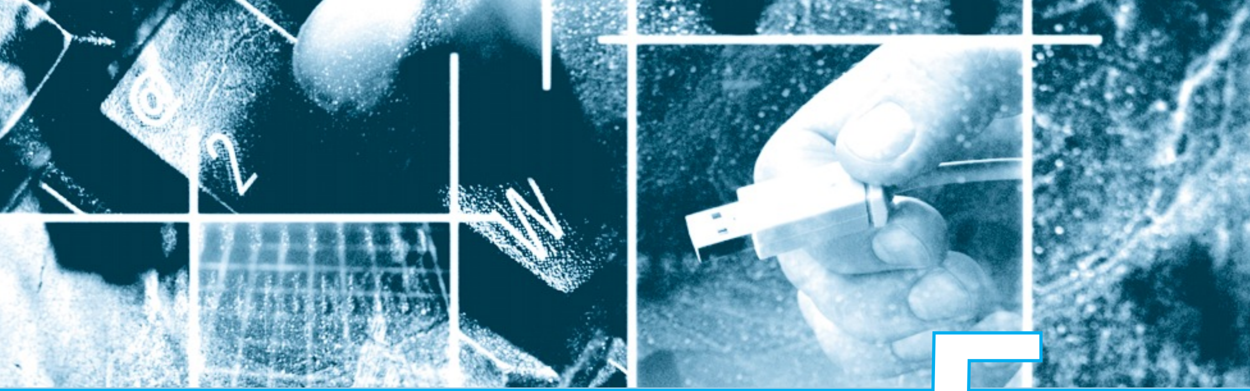
Note the following relationships:

$$\begin{aligned} +30 \text{ dBm} &= 0 \text{ dBW} \\ 0 \text{ dBm} &= -30 \text{ dBW} \end{aligned}$$

A unit in common use in cable television and broadband LAN applications is the **dBmV (decibel-millivolt)**. This is an absolute unit with 0 dBmV equivalent to 1 mV. Thus

$$\text{Voltage}_{\text{dBmV}} = 20 \log \frac{\text{Voltage}_{\text{mV}}}{1 \text{ mV}}$$

In this case, the voltage levels are assumed to be across a 75-ohm resistance.



CHAPTER 5

Signal Encoding Techniques

Even the natives have difficulty mastering this peculiar vocabulary.

—*The Golden Bough*, Sir James George Frazer

KEY POINTS

- Both analog and digital information can be encoded as either analog or digital signals. The particular encoding that is chosen depends on the specific requirements to be met and the media and communications facilities available.
- **Digital data, digital signals:** The simplest form of digital encoding of digital data is to assign one voltage level to binary one and another to binary zero. More complex encoding schemes are used to improve performance, by altering the spectrum of the signal and providing synchronization capability.
- **Digital data, analog signal:** A modem converts digital data to an analog signal so that it can be transmitted over an analog line. The basic techniques are amplitude shift keying (ASK), frequency shift keying (FSK), and phase shift keying (PSK). All involve altering one or more characteristics of a carrier frequency to represent binary data.
- **Analog data, digital signals:** Analog data, such as voice and video, are often digitized to be able to use digital transmission facilities. The simplest technique is pulse code modulation (PCM), which involves sampling the analog data periodically and quantizing the samples.
- **Analog data, analog signals:** Analog data are modulated by a carrier frequency to produce an analog signal in a different frequency band, which can be utilized on an analog transmission system. The basic techniques are amplitude modulation (AM), frequency modulation (FM), and phase modulation (PM).

In Chapter 3 a distinction was made between analog and digital data and analog and digital signals. Figure 3.14 suggested that either form of data could be encoded into either form of signal.

Figure 5.1 is another depiction that emphasizes the process involved. For **digital signaling**, a data source $g(t)$, which may be either digital or analog, is encoded into a digital signal $x(t)$. The actual form of $x(t)$ depends on the encoding technique and is chosen to optimize use of the transmission medium. For example, the encoding may be chosen to conserve bandwidth or to minimize errors.

The basis for **analog signaling** is a continuous constant-frequency signal known as the **carrier signal**. The frequency of the carrier signal is chosen to be compatible with the transmission medium being used. Data may be transmitted using a carrier signal by modulation. **Modulation** is the process of encoding

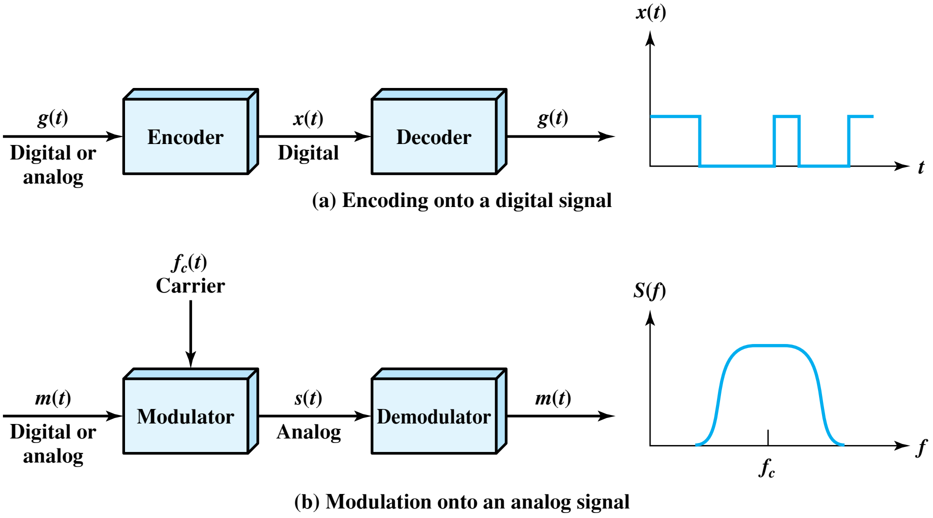


Figure 5.1 Encoding and Modulation Techniques

source data onto a carrier signal with frequency f_c. All modulation techniques involve operation on one or more of the three fundamental frequency domain parameters: amplitude, frequency, and phase.

The input signal $m(t)$ may be analog or digital and is called the modulating signal or **baseband signal**. The result of modulating the carrier signal is called the modulated signal $s(t)$. As Figure 5.1b indicates, $s(t)$ is a bandlimited (bandpass) signal. The location of the bandwidth on the spectrum is related to f_c and is often centered on f_c. Again, the actual form of the encoding is chosen to optimize some characteristic of the transmission.

Each of the four possible combinations depicted in Figure 5.1 is in widespread use. The reasons for choosing a particular combination for any given communication task vary. We list here some representative reasons:

- **Digital data, digital signal:** In general, the equipment for encoding digital data into a digital signal is less complex and less expensive than digital-to-analog modulation equipment.
- **Analog data, digital signal:** Conversion of analog data to digital form permits the use of modern digital transmission and switching equipment. The advantages of the digital approach were outlined in Section 3.2.
- **Digital data, analog signal:** Some transmission media, such as optical fiber and unguided media, will only propagate analog signals.
- **Analog data, analog signal:** Analog data in electrical form can be transmitted as baseband signals easily and cheaply. This is done with voice transmission over voice-grade lines. One common use of modulation is to shift the bandwidth of a baseband signal to another portion of the spectrum. In this way multiple signals, each at a different position on the

spectrum, can share the same transmission medium. This is known as frequency division multiplexing.

We now examine the techniques involved in each of these four combinations.

5.1 DIGITAL DATA, DIGITAL SIGNALS

A digital signal is a sequence of discrete, discontinuous voltage pulses. Each pulse is a signal element. Binary data are transmitted by encoding each data bit into signal elements. In the simplest case, there is a one-to-one correspondence between bits and signal elements. An example is shown in Figure 3.16, in which binary 1 is represented by a lower voltage level and binary 0 by a higher voltage level. We show in this section that a variety of other encoding schemes are also used.

First, we define some terms. If the signal elements all have the same algebraic sign, that is, all positive or negative, then the signal is **unipolar**. In **polar** signaling, one logic state is represented by a positive voltage level, and the other by a negative voltage level. The **data signaling rate**, or just **data rate**, of a signal is the rate, in bits per second, that data are transmitted. The duration or length of a bit is the amount of time it takes for the transmitter to emit the bit; for a data rate R, the bit duration is $1/R$. The **modulation rate**, in contrast, is the rate at which the signal level is changed. This will depend on the nature of the digital encoding, as explained later. The modulation rate is expressed in baud, which means signal elements per second. Finally, the terms mark and space, for historical reasons, refer to the binary digits 1 and 0, respectively. Table 5.1 summarizes key terms; these should be clearer when we see an example later in this section.

The tasks involved in interpreting digital signals at the receiver can be summarized by again referring to Figure 3.16. First, the receiver must know the timing of each bit. That is, the receiver must know with some accuracy when a bit begins and ends. Second, the receiver must determine whether the signal level for each bit position is high (0) or low (1). In Figure 3.16, these tasks are performed by sampling each bit position in the middle of the interval and comparing the value to a threshold. Because of noise and other impairments, there will be errors, as shown.

What factors determine how successful the receiver will be in interpreting the incoming signal? We saw in Chapter 3 that three factors are important: the

Table 5.1 Key Data Transmission Terms

Term	Units	Definition
Data element	Bits	A single binary one or zero
Data rate	Bits per second (bps)	The rate at which data elements are transmitted
Signal element	Digital: a voltage pulse of constant amplitude Analog: a pulse of constant frequency, phase, and amplitude	That part of a signal that occupies the shortest interval of a signaling code
Signaling rate or modulation rate	Signal elements per second (baud)	The rate at which signal elements are transmitted

Table 5.2 Definition of Digital Signal Encoding Formats

Nonreturn to Zero-Level (NRZ-L)
0 = high level
1 = low level

Nonreturn to Zero Inverted (NRZI)
0 = no transition at beginning of interval (one bit time)
1 = transition at beginning of interval

Bipolar-AMI
0 = no line signal
1 = positive or negative level, alternating for successive ones

Pseudoternary
0 = positive or negative level, alternating for successive zeros
1 = no line signal

Manchester
0 = transition from high to low in middle of interval
1 = transition from low to high in middle of interval

Differential Manchester
Always a transition in middle of interval
0 = transition at beginning of interval
1 = no transition at beginning of interval

B8ZS
Same as bipolar AMI, except that any string of eight zeros is replaced by a string with two code violations

HDB3
Same as bipolar AMI, except that any string of four zeros is replaced by a string with one code violation

signal-to-noise ratio, the data rate, and the bandwidth. With other factors held constant, the following statements are true:

- An increase in data rate increases bit error rate (BER).[1]
- An increase in SNR decreases bit error rate.
- An increase in bandwidth allows an increase in data rate.

There is another factor that can be used to improve performance, and that is the encoding scheme. The encoding scheme is simply the mapping from data bits to signal elements. A variety of approaches have been tried. In what follows, we describe some of the more common ones; they are defined in Table 5.2 and depicted in Figure 5.2.

Before describing these techniques, let us consider the following ways of evaluating or comparing the various techniques.

[1] The BER is the most common measure of error performance on a data circuit and is defined as the probability that a bit is received in error. It is also called the *bit error ratio*. This latter term is clearer, because the term *rate* typically refers to some quantity that varies with time. Unfortunately, most books and standards documents refer to the R in BER as *rate*.

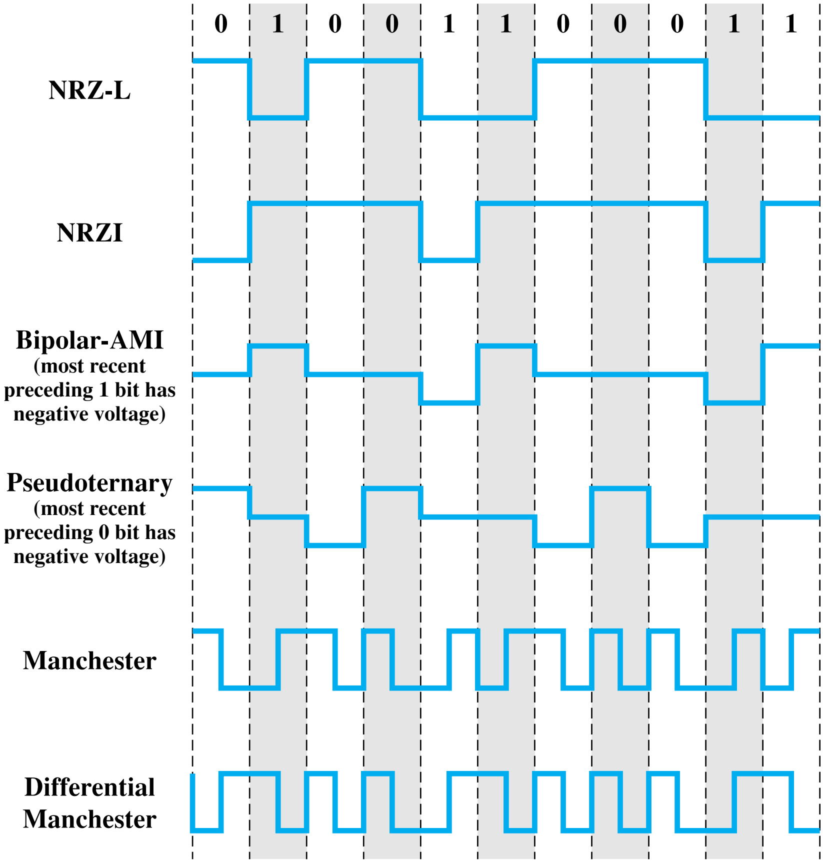

0
1
0
0
1
1
0
0
0
1
1
NRZ-L
NRZI
Bipolar-AMI
(most recent preceding 1 bit has negative voltage)
Pseudoternary
(most recent preceding 0 bit has negative voltage)
Manchester
Differential Manchester

Figure 5.2 Digital Signal Encoding Formats

- **Signal spectrum:** Several aspects of the signal spectrum are important. A lack of high-frequency components means that less bandwidth is required for transmission. In addition, lack of a direct-current (dc) component is also desirable. With a dc component to the signal, there must be direct physical attachment of transmission components. With no dc component, ac coupling via transformer is possible; this provides excellent electrical isolation, reducing interference. Finally, the magnitude of the effects of signal distortion and interference depend on the spectral properties of the transmitted signal. In practice, it usually happens that the transmission characteristics of a channel are worse near the band edges. Therefore, a good signal design should concentrate the transmitted power in the middle of the transmission bandwidth. In such a case, a smaller distortion should be present in the received signal. To meet this objective, codes can be designed with the aim of shaping the spectrum of the transmitted signal.
- **Clocking:** We mentioned the need to determine the beginning and end of each bit position. This is no easy task. One rather expensive approach is to provide

a separate clock lead to synchronize the transmitter and receiver. The alternative is to provide some synchronization mechanism that is based on the transmitted signal. This can be achieved with suitable encoding, as explained subsequently.

- **Error detection:** We will discuss various error-detection techniques in Chapter 6 and show that these are the responsibility of a layer of logic above the signaling level that is known as data link control. However, it is useful to have some error detection capability built into the physical signaling encoding scheme. This permits errors to be detected more quickly.
- **Signal interference and noise immunity:** Certain codes exhibit superior performance in the presence of noise. Performance is usually expressed in terms of a BER.
- **Cost and complexity:** Although digital logic continues to drop in price, this factor should not be ignored. In particular, the higher the signaling rate to achieve a given data rate, the greater the cost. We shall see that some codes require a signaling rate that is greater than the actual data rate.

We now turn to a discussion of various techniques.

Nonreturn to Zero (NRZ)

The most common, and easiest, way to transmit digital signals is to use two different voltage levels for the two binary digits. Codes that follow this strategy share the property that the voltage level is constant during a bit interval; there is no transition (no return to a zero voltage level). For example, the absence of voltage can be used to represent binary 0, with a constant positive voltage used to represent binary 1. More commonly, a negative voltage represents one binary value and a positive voltage represents the other. This latter code, known as **Nonreturn to Zero-Level** (NRZ-L), is illustrated[2] in Figure 5.2. NRZ-L is typically the code used to generate or interpret digital data by terminals and other devices. If a different code is to be used for transmission, it is generated from an NRZ-L signal by the transmission system [in terms of Figure 5.1, NRZ-L is $g(t)$ and the encoded signal is $x(t)$].

A variation of NRZ is known as **NRZI** (Nonreturn to Zero, invert on ones). As with NRZ-L, NRZI maintains a constant voltage pulse for the duration of a bit time. The data themselves are encoded as the presence or absence of a signal transition at the beginning of the bit time. A transition (low to high or high to low) at the beginning of a bit time denotes a binary 1 for that bit time; no transition indicates a binary 0.

NRZI is an example of **differential encoding**. In differential encoding, the information to be transmitted is represented in terms of the changes between successive signal elements rather than the signal elements themselves. The encoding of the current bit is determined as follows: If the current bit is a binary 0, then the

[2] In this figure, a negative voltage is equated with binary 1 and a positive voltage with binary 0. This is the opposite of the definition used in virtually all other textbooks. The definition here conforms to the use of NRZ-L in data communications interfaces and the standards that govern those interfaces.

current bit is encoded with the same signal as the preceding bit; if the current bit is a binary 1, then the current bit is encoded with a different signal than the preceding bit. One benefit of differential encoding is that it may be more reliable to detect a transition in the presence of noise than to compare a value to a threshold. Another benefit is that with a complex transmission layout, it is easy to lose the sense of the polarity of the signal. For example, on a multidrop twisted-pair line, if the leads from an attached device to the twisted pair are accidentally inverted, all 1s and 0s for NRZ-L will be inverted. This does not happen with differential encoding.

The NRZ codes are the easiest to engineer and, in addition, make efficient use of bandwidth. This latter property is illustrated in Figure 5.3, which compares the spectral density of various encoding schemes. In the figure, frequency is normalized to the data rate. Most of the energy in NRZ and NRZI signals is between dc and half the bit rate. For example, if an NRZ code is used to generate a signal with data rate of 9600 bps, most of the energy in the signal is concentrated between dc and 4800 Hz.

The main limitations of NRZ signals are the presence of a dc component and the lack of synchronization capability. To picture the latter problem, consider that with a long string of 1s or 0s for NRZ-L or a long string of 0s for NRZI, the output is a constant voltage over a long period of time. Under these circumstances, any drift between the clocks of transmitter and receiver will result in loss of synchronization between the two.

Because of their simplicity and relatively low frequency response characteristics, NRZ codes are commonly used for digital magnetic recording. However, their limitations make these codes unattractive for signal transmission applications.

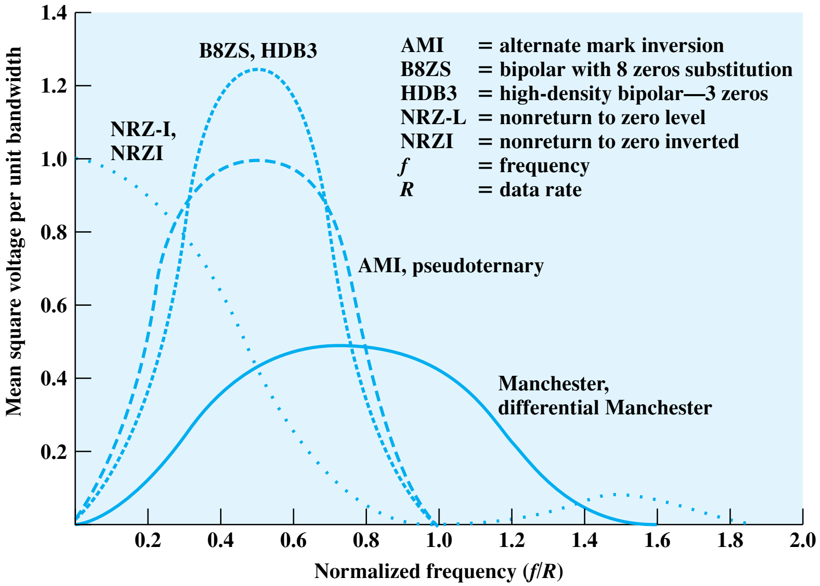


Figure 5.3 Spectral Density of Various Signal Encoding Schemes

Multilevel Binary

A category of encoding techniques known as multilevel binary addresses some of the deficiencies of the NRZ codes. These codes use more than two signal levels. Two examples of this scheme are illustrated in Figure 5.2, bipolar-AMI (alternate mark inversion) and pseudoternary.[3]

In the case of the **bipolar-AMI** scheme, a binary 0 is represented by no line signal, and a binary 1 is represented by a positive or negative pulse. The binary 1 pulses must alternate in polarity. There are several advantages to this approach. First, there will be no loss of synchronization if a long string of 1s occurs. Each 1 introduces a transition, and the receiver can resynchronize on that transition. A long string of 0s would still be a problem. Second, because the 1 signals alternate in voltage from positive to negative, there is no net dc component. Also, the bandwidth of the resulting signal is considerably less than the bandwidth for NRZ (Figure 5.3). Finally, the pulse alternation property provides a simple means of error detection. Any isolated error, whether it deletes a pulse or adds a pulse, causes a violation of this property.

The comments of the previous paragraph also apply to **pseudoternary**. In this case, it is the binary 1 that is represented by the absence of a line signal, and the binary 0 by alternating positive and negative pulses. There is no particular advantage of one technique versus the other, and each is the basis of some applications.

Although a degree of synchronization is provided with these codes, a long string of 0s in the case of AMI or 1s in the case of pseudoternary still presents a problem. Several techniques have been used to address this deficiency. One approach is to insert additional bits that force transitions. This technique is used in ISDN (integrated services digital network) for relatively low data rate transmission. Of course, at a high data rate, this scheme is expensive, because it results in an increase in an already high signal transmission rate. To deal with this problem at high data rates, a technique that involves scrambling the data is used. We examine two examples of this technique later in this section.

Thus, with suitable modification, multilevel binary schemes overcome the problems of NRZ codes. Of course, as with any engineering design decision, there is a tradeoff. With multilevel binary coding, the line signal may take on one of three levels, but each signal element, which could represent $\log_2 3 = 1.58$ bits of information, bears only one bit of information. Thus multilevel binary is not as efficient as NRZ coding. Another way to state this is that the receiver of multilevel binary signals has to distinguish between three levels $(+A, -A, 0)$ instead of just two levels in the signaling formats previously discussed. Because of this, the multilevel binary signal requires approximately 3 dB more signal power than a two-valued signal for the same probability of bit error. This is illustrated in Figure 5.4. Put another way, the bit error rate for NRZ codes, at a given signal-to-noise ratio, is significantly less than that for multilevel binary.

[3] These terms are not used consistently in the literature. In some books, these two terms are used for different encoding schemes than those defined here, and a variety of terms have been used for the two schemes illustrated in Figure 5.2. The nomenclature used here corresponds to the usage in various ITU-T standards documents.

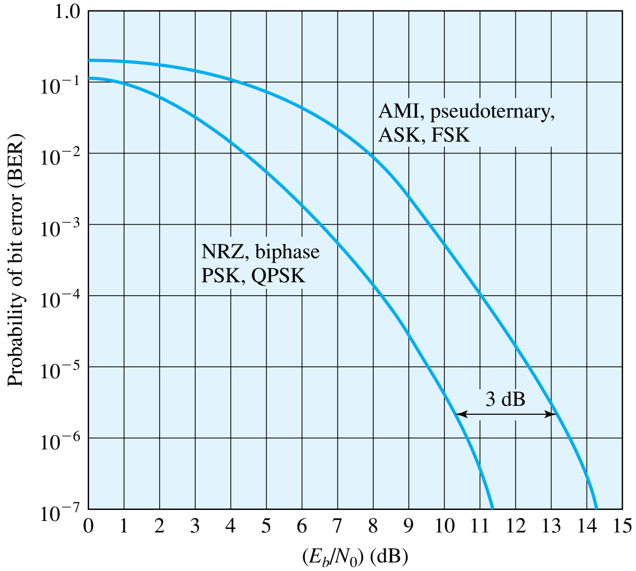


Figure 5.4 Theoretical Bit Error Rate for Various Encoding Schemes

Biphase

There is another set of coding techniques, grouped under the term *biphase*, that overcomes the limitations of NRZ codes. Two of these techniques, Manchester and differential Manchester, are in common use.

In the **Manchester** code, there is a transition at the middle of each bit period. The midbit transition serves as a clocking mechanism and also as data: a low-to-high transition represents a 1, and a high-to-low transition represents a 0.[4] In **differential Manchester**, the midbit transition is used only to provide clocking. The encoding of a 0 is represented by the presence of a transition at the beginning of a bit period, and a 1 is represented by the absence of a transition at the beginning of a bit period. Differential Manchester has the added advantage of employing differential encoding.

All of the biphase techniques require at least one transition per bit time and may have as many as two transitions. Thus, the maximum modulation rate is twice that for NRZ; this means that the bandwidth required is correspondingly greater. On the other hand, the biphase schemes have several advantages:

- **Synchronization:** Because there is a predictable transition during each bit time, the receiver can synchronize on that transition. For this reason, the biphase codes are known as self-clocking codes.
- **No dc component:** Biphase codes have no dc component, yielding the benefits described earlier.

[4] The definition of Manchester presented here is the opposite of that used in a number of respectable textbooks, in which a low-to-high transition represents a binary 0 and a high-to-low transition represents a binary 1. Here, we conform to industry practice and to the definition used in the various LAN standards, such as IEEE 802.3.

- **Error detection:** The absence of an expected transition can be used to detect errors. Noise on the line would have to invert both the signal before and after the expected transition to cause an undetected error.

As can be seen from Figure 5.3, the bandwidth for biphase codes is reasonably narrow and contains no dc component. However, it is wider than the bandwidth for the multilevel binary codes.

Biphase codes are popular techniques for data transmission. The more common Manchester code has been specified for the IEEE 802.3 (Ethernet) standard for baseband coaxial cable and twisted-pair bus LANs. Differential Manchester has been specified for the IEEE 802.5 token ring LAN, using shielded twisted pair.

Modulation Rate

When signal-encoding techniques are used, a distinction needs to be made between data rate (expressed in bits per second) and modulation rate (expressed in baud). The data rate, or bit rate, is $1/T_b$, where $T_b =$ bit duration. The modulation rate is the rate at which signal elements are generated. Consider, for example, Manchester encoding. The minimum size signal element is a pulse of one-half the duration of a bit interval. For a string of all binary zeroes or all binary ones, a continuous stream of such pulses is generated. Hence the maximum modulation rate for Manchester is $2/T_b$. This situation is illustrated in Figure 5.5, which shows the transmission of a stream of binary 1s at a data rate of 1 Mbps using NRZI and Manchester. In general,

$$D = \frac{R}{L} = \frac{R}{\log_2 M} \tag{5.1}$$

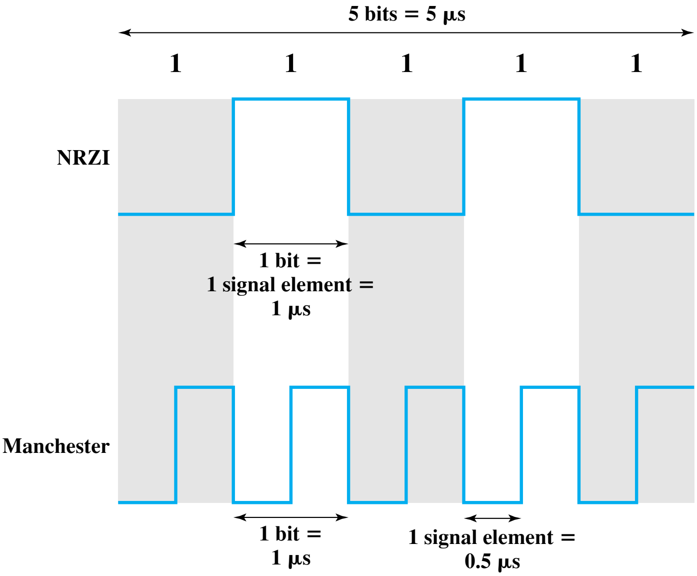


Figure 5.5 A Stream of Binary Ones at 1 Mbps

Table 5.3 Normalized Signal Transition Rate of Various Digital Signal Encoding Schemes

	Minimum	101010 . . .	Maximum
NRZ-L	0 (all 0s or 1s)	1.0	1.0
NRZI	0 (all 0s)	0.5	1.0 (all 1s)
Bipolar-AMI	0 (all 0s)	1.0	1.0
Pseudoternary	0 (all 1s)	1.0	1.0
Manchester	1.0 (1010 . . .)	1.0	2.0 (all 0s or 1s)
Differential Manchester	1.0 (all 1s)	1.5	2.0 (all 0s)

where

D = modulation rate, baud
R = data rate, bps
M = number of different signal elements $= 2^L$
L = number of bits per signal element

One way of characterizing the modulation rate is to determine the average number of transitions that occur per bit time. In general, this will depend on the exact sequence of bits being transmitted. Table 5.3 compares transition rates for various techniques. It indicates the signal transition rate in the case of a data stream of alternating 1s and 0s, and for the data stream that produces the minimum and maximum modulation rate.

Scrambling Techniques

Although the biphase techniques have achieved widespread use in local area network applications at relatively high data rates (up to 10 Mbps), they have not been widely used in long-distance applications. The principal reason for this is that they require a high signaling rate relative to the data rate. This sort of inefficiency is more costly in a long-distance application.

Another approach is to make use of some sort of scrambling scheme. The idea behind this approach is simple: Sequences that would result in a constant voltage level on the line are replaced by filling sequences that will provide sufficient transitions for the receiver's clock to maintain synchronization. The filling sequence must be recognized by the receiver and replaced with the original data sequence. The filling sequence is the same length as the original sequence, so there is no data rate penalty. The design goals for this approach can be summarized as follows:

- No dc component
- No long sequences of zero-level line signals
- No reduction in data rate
- Error-detection capability

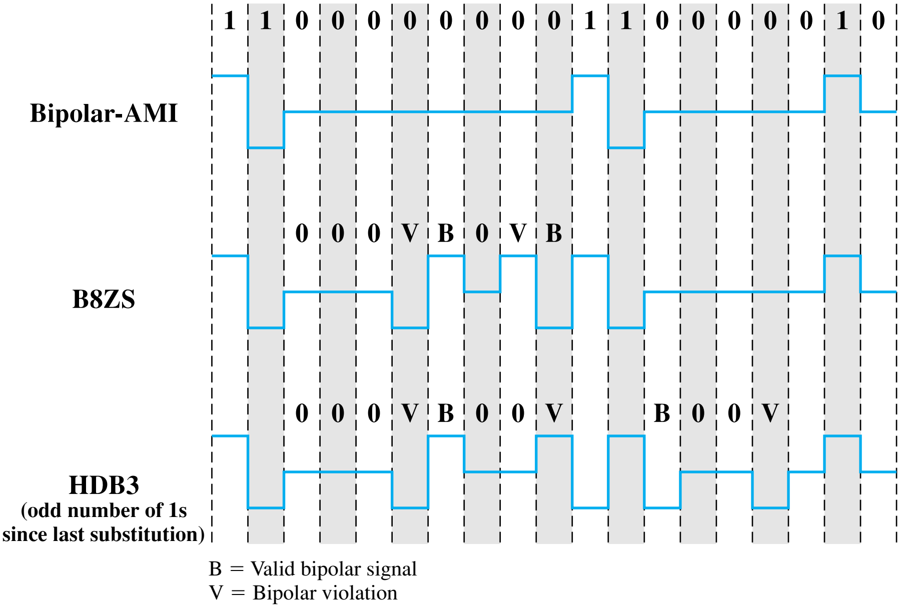


Figure 5.6 Encoding Rules for B8ZS and HDB3

Two techniques are commonly used in long-distance transmission services; these are illustrated in Figure 5.6.

A coding scheme that is commonly used in North America is known as **bipolar with 8-zeros substitution (B8ZS)**. The coding scheme is based on a bipolar-AMI. We have seen that the drawback of the AMI code is that a long string of zeros may result in loss of synchronization. To overcome this problem, the encoding is amended with the following rules:

- If an octet of all zeros occurs and the last voltage pulse preceding this octet was positive, then the eight zeros of the octet are encoded as 000+−0−+.
- If an octet of all zeros occurs and the last voltage pulse preceding this octet was negative, then the eight zeros of the octet are encoded as 000−+0+−.

This technique forces two code violations (signal patterns not allowed in AMI) of the AMI code, an event unlikely to be caused by noise or other transmission impairment. The receiver recognizes the pattern and interprets the octet as consisting of all zeros.

A coding scheme that is commonly used in Europe and Japan is known as the **high-density bipolar-3 zeros (HDB3)** code (Table 5.4). As before, it is based on the use of AMI encoding. In this case, the scheme replaces strings of four zeros with sequences containing one or two pulses. In each case, the fourth zero is replaced with a code violation. In addition, a rule is needed to ensure that successive violations are of alternate polarity so that no dc component is introduced. Thus, if the last violation was positive, this violation must be negative and vice versa. Table 5.4 shows that this condition is tested for by determining (1) whether the number of

Table 5.4 HDB3 Substitution Rules

	Number of Bipolar Pulses (ones) since Last Substitution	
Polarity of Preceding Pulse	**Odd**	**Even**
−	0 0 0 −	+ 0 0 +
+	0 0 0 +	− 0 0 −

pulses since the last violation is even or odd and (2) the polarity of the last pulse before the occurrence of the four zeros.

Figure 5.3 shows the spectral properties of these two codes. As can be seen, neither has a dc component. Most of the energy is concentrated in a relatively sharp spectrum around a frequency equal to one-half the data rate. Thus, these codes are well suited to high data rate transmission.

5.2 DIGITAL DATA, ANALOG SIGNALS

We turn now to the case of transmitting digital data using analog signals. The most familiar use of this transformation is for transmitting digital data through the public telephone network. The telephone network was designed to receive, switch, and transmit analog signals in the voice-frequency range of about 300 to 3400 Hz. It is not at present suitable for handling digital signals from the subscriber locations (although this is beginning to change). Thus digital devices are attached to the network via a modem (modulator-demodulator), which converts digital data to analog signals, and vice versa.

For the telephone network, modems are used that produce signals in the voice-frequency range. The same basic techniques are used for modems that produce signals at higher frequencies (e.g., microwave). This section introduces these techniques and provides a brief discussion of the performance characteristics of the alternative approaches.

We mentioned that modulation involves operation on one or more of the three characteristics of a carrier signal: amplitude, frequency, and phase. Accordingly, there are three basic encoding or modulation techniques for transforming digital data into analog signals, as illustrated in Figure 5.7: amplitude shift keying (ASK), frequency shift keying (FSK), and phase shift keying (PSK). In all these cases, the resulting signal occupies a bandwidth centered on the carrier frequency.

Amplitude Shift Keying

In ASK, the two binary values are represented by two different amplitudes of the carrier frequency. Commonly, one of the amplitudes is zero; that is, one binary digit is represented by the presence, at constant amplitude, of the carrier, the other by the absence of the carrier (Figure 5.7a). The resulting transmitted signal for one bit time is

$$\textbf{ASK} \qquad s(t) = \begin{cases} A\cos(2\pi f_c t) & \text{binary 1} \\ 0 & \text{binary 0} \end{cases} \tag{5.2}$$

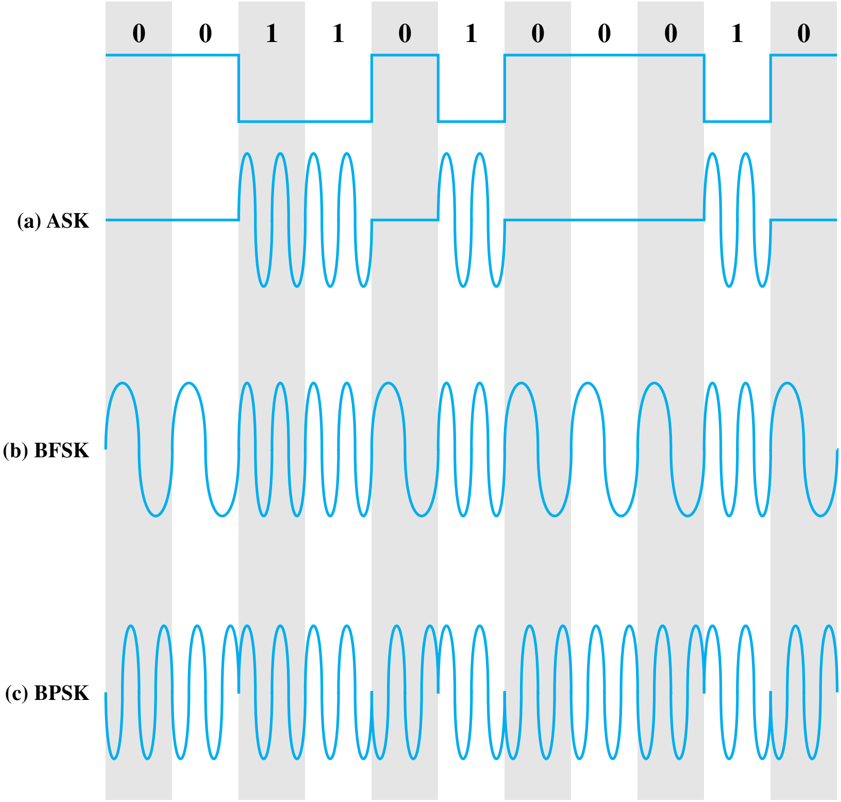


Figure 5.7 Modulation of Analog Signals for Digital Data

where the carrier signal is $A\cos(2\pi f_c t)$. ASK is susceptible to sudden gain changes and is a rather inefficient modulation technique. On voice-grade lines, it is typically used only up to 1200 bps.

The ASK technique is used to transmit digital data over optical fiber. For LED (light-emitting diode) transmitters, Equation (5.2) is valid. That is, one signal element is represented by a light pulse while the other signal element is represented by the absence of light. Laser transmitters normally have a fixed "bias" current that causes the device to emit a low light level. This low level represents one signal element, while a higher-amplitude lightwave represents another signal element.

Frequency Shift Keying

The most common form of FSK is binary FSK (BFSK), in which the two binary values are represented by two different frequencies near the carrier frequency (Figure 5.7b). The resulting transmitted signal for one bit time is

$$\textbf{BFSK} \qquad s(t) = \begin{cases} A\cos(2\pi f_1 t) & \text{binary 1} \\ A\cos(2\pi f_2 t) & \text{binary 0} \end{cases} \tag{5.3}$$

where f_1 and f_2 are typically offset from the carrier frequency f_c by equal but opposite amounts.

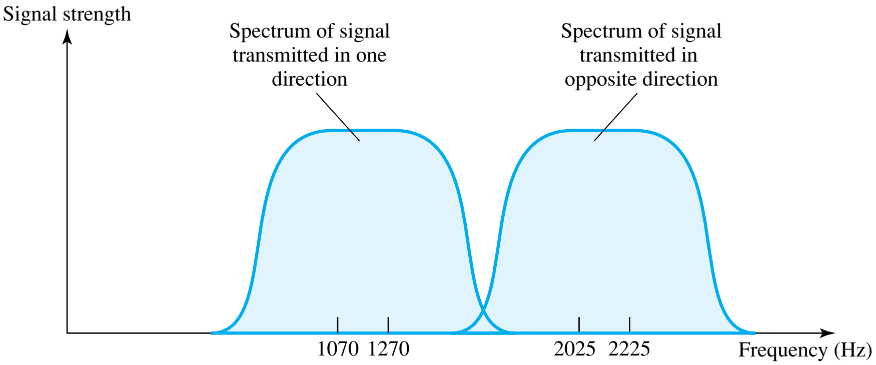


Figure 5.8 Full-Duplex FSK Transmission on a Voice-Grade Line

Figure 5.8 shows an example of the use of BFSK for full-duplex operation over a voice-grade line. The figure is a specification for the Bell System 108 series modems. Recall that a voice-grade line will pass frequencies in the approximate range 300 to 3400 Hz and that *full duplex* means that signals are transmitted in both directions at the same time. To achieve full-duplex transmission, this bandwidth is split. In one direction (transmit or receive), the frequencies used to represent 1 and 0 are centered on 1170 Hz, with a shift of 100 Hz on either side. The effect of alternating between those two frequencies is to produce a signal whose spectrum is indicated as the shaded area on the left in Figure 5.8. Similarly, for the other direction (receive or transmit) the modem uses frequencies shifted 100 Hz to each side of a center frequency of 2125 Hz. This signal is indicated by the shaded area on the right in Figure 5.8. Note that there is little overlap and thus little interference.

BFSK is less susceptible to error than ASK. On voice-grade lines, it is typically used up to 1200 bps. It is also commonly used for high-frequency (3 to 30 MHz) radio transmission. It can also be used at even higher frequencies on local area networks that use coaxial cable.

A signal that is more bandwidth efficient, but also more susceptible to error, is multiple FSK (MFSK), in which more than two frequencies are used. In this case each signaling element represents more than one bit. The transmitted MFSK signal for one signal element time can be defined as follows:

$$\textbf{MFSK} \qquad s_i(t) = A\cos 2\pi f_i t, \qquad 1 \le i \le M \tag{5.4}$$

where

$$\begin{aligned} f_i &= f_c + (2i - 1 - M)f_d \\ f_c &= \text{the carrier frequency} \\ f_d &= \text{the difference frequency} \\ M &= \text{number of different signal elements} = 2^L \\ L &= \text{number of bits per signal element} \end{aligned}$$

To match the data rate of the input bit stream, each output signal element is held for a period of $T_s = LT$ seconds, where T is the bit period (data rate $= 1/T$). Thus, one signal element, which is a constant-frequency tone, encodes L bits. The total bandwidth required is $2Mf_d$. It can be shown that the minimum frequency separation required is $2f_d = 1/T_s$. Therefore, the modulator requires a bandwidth of $W_d = 2Mf_d = M/T_s$.

EXAMPLE 5.1 With $f_c = 250$ kHz, $f_d = 25$ kHz, and $M = 8$ ($L = 3$ bits), we have the following frequency assignments for each of the eight possible 3-bit data combinations:

$f_1 = 75$ kHz	000	$f_2 = 125$ kHz	001
$f_3 = 175$ kHz	010	$f_4 = 225$ kHz	011
$f_5 = 275$ kHz	100	$f_6 = 325$ kHz	101
$f_7 = 375$ kHz	110	$f_8 = 425$ kHz	111

This scheme can support a data rate of $1/T = 2Lf_d = 150$ kbps.

EXAMPLE 5.2 Figure 5.9 shows an example of MFSK with $M = 4$. An input bit stream of 20 bits is encoded 2 bits at a time, with each of the four possible 2-bit combinations transmitted as a different frequency. The display in the figure shows the frequency transmitted (y-axis) as a function of time (x-axis). Each column represents a time unit T_s in which a single 2-bit signal element is transmitted. The shaded rectangle in the column indicates the frequency transmitted during that time unit.

Phase Shift Keying

In PSK, the phase of the carrier signal is shifted to represent data.

Two-Level PSK The simplest scheme uses two phases to represent the two binary digits (Figure 5.7c) and is known as binary phase shift keying. The resulting transmitted signal for one bit time is

$$\textbf{BPSK} \qquad s(t) = \begin{cases} A\cos(2\pi f_c t) \\ A\cos(2\pi f_c t + \pi) \end{cases} = \begin{cases} A\cos(2\pi f_c t) & \text{binary 1} \\ -A\cos(2\pi f_c t) & \text{binary 0} \end{cases} \tag{5.5}$$

Because a phase shift of 180° (π) is equivalent to flipping the sine wave or multiplying it by -1, the rightmost expressions in Equation (5.5) can be used. This

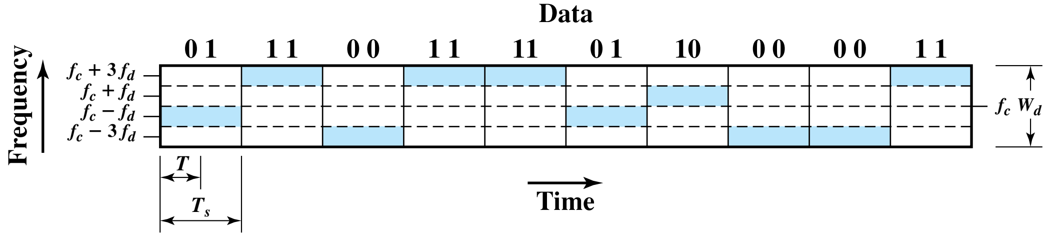


Figure 5.9 MFSK Frequency Use ($M = 4$)

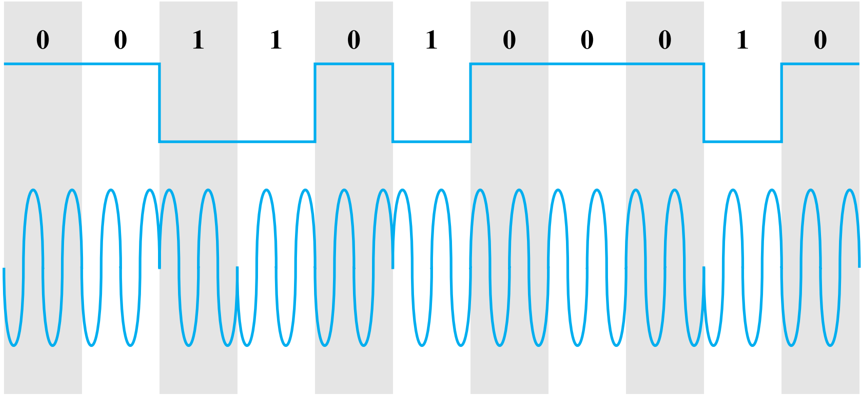


Figure 5.10 Differential Phase-Shift Keying (DPSK)

leads to a convenient formulation. If we have a bit stream, and we define $d(t)$ as the discrete function that takes on the value of $+1$ for one bit time if the corresponding bit in the bit stream is 1 and the value of -1 for one bit time if the corresponding bit in the bit stream is 0, then we can define the transmitted signal as

$$\textbf{BPSK} \qquad s_d(t) = A\, d(t)\cos(2\pi f_c t) \tag{5.6}$$

An alternative form of two-level PSK is differential PSK (DPSK). Figure 5.10 shows an example. In this scheme, a binary 0 is represented by sending a signal burst of the same phase as the previous signal burst sent. A binary 1 is represented by sending a signal burst of opposite phase to the preceding one. This term *differential* refers to the fact that the phase shift is with reference to the previous bit transmitted rather than to some constant reference signal. In differential encoding, the information to be transmitted is represented in terms of the changes between successive data symbols rather than the signal elements themselves. DPSK avoids the requirement for an accurate local oscillator phase at the receiver that is matched with the transmitter. As long as the preceding phase is received correctly, the phase reference is accurate.

Four-Level PSK More efficient use of bandwidth can be achieved if each signaling element represents more than one bit. For example, instead of a phase shift of 180°, as allowed in BPSK, a common encoding technique, known as quadrature phase shift keying (QPSK), uses phase shifts separated by multiples of $\pi/2$ (90°).

$$\textbf{QPSK} \qquad s(t) = \begin{cases} A\cos\left(2\pi f_c t + \dfrac{\pi}{4}\right) & 11 \\ A\cos\left(2\pi f_c t + \dfrac{3\pi}{4}\right) & 01 \\ A\cos\left(2\pi f_c t - \dfrac{3\pi}{4}\right) & 00 \\ A\cos\left(2\pi f_c t - \dfrac{\pi}{4}\right) & 10 \end{cases} \tag{5.7}$$

Thus each signal element represents two bits rather than one.

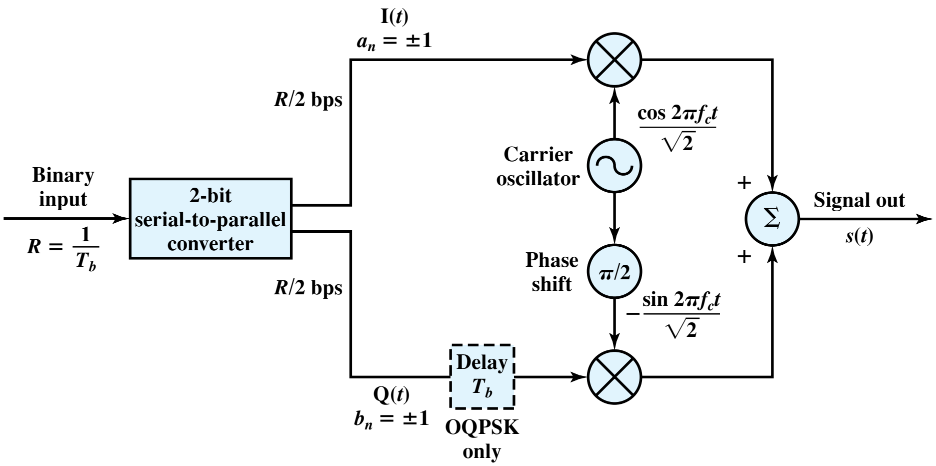


Figure 5.11 QPSK and OQPSK Modulators

Figure 5.11 shows the QPSK modulation scheme in general terms. The input is a stream of binary digits with a data rate of $R = 1/T_b$, where T_b is the width of each bit. This stream is converted into two separate bit streams of $R/2$ bps each, by taking alternate bits for the two streams. The two data streams are referred to as the I (in-phase) and Q (quadrature phase) streams. In the diagram, the upper stream is modulated on a carrier of frequency f_c by multiplying the bit stream by the carrier. For convenience of modulator structure we map binary 1 to $\sqrt{1/2}$ and binary 0 to $-\sqrt{1/2}$. Thus, a binary 1 is represented by a scaled version of the carrier wave and a binary 0 is represented by a scaled version of the negative of the carrier wave, both at a constant amplitude. This same carrier wave is shifted by 90° and used for modulation of the lower binary stream. The two modulated signals are then added together and transmitted. The transmitted signal can be expressed as follows:

$$\textbf{QPSK} \qquad s(t) = \frac{1}{\sqrt{2}} I(t) \cos 2\pi f_c t - \frac{1}{\sqrt{2}} Q(t) \sin 2\pi f_c t$$

Figure 5.12 shows an example of QPSK coding. Each of the two modulated streams is a BPSK signal at half the data rate of the original bit stream. Thus, the combined signals have a symbol rate that is half the input bit rate. Note that from one symbol time to the next, a phase change of as much as 180° (π) is possible.

Figure 5.11 also shows a variation of QPSK known as offset QPSK (OQPSK), or orthogonal QPSK. The difference is that a delay of one bit time is introduced in the Q stream, resulting in the following signal:

$$s(t) = \frac{1}{\sqrt{2}} I(t) \cos 2\pi f_c t - \frac{1}{\sqrt{2}} Q(t - T_b) \sin 2\pi f_c t$$

Because OQPSK differs from QPSK only by the delay in the Q stream, its spectral characteristics and bit error performance are the same as that of QPSK.

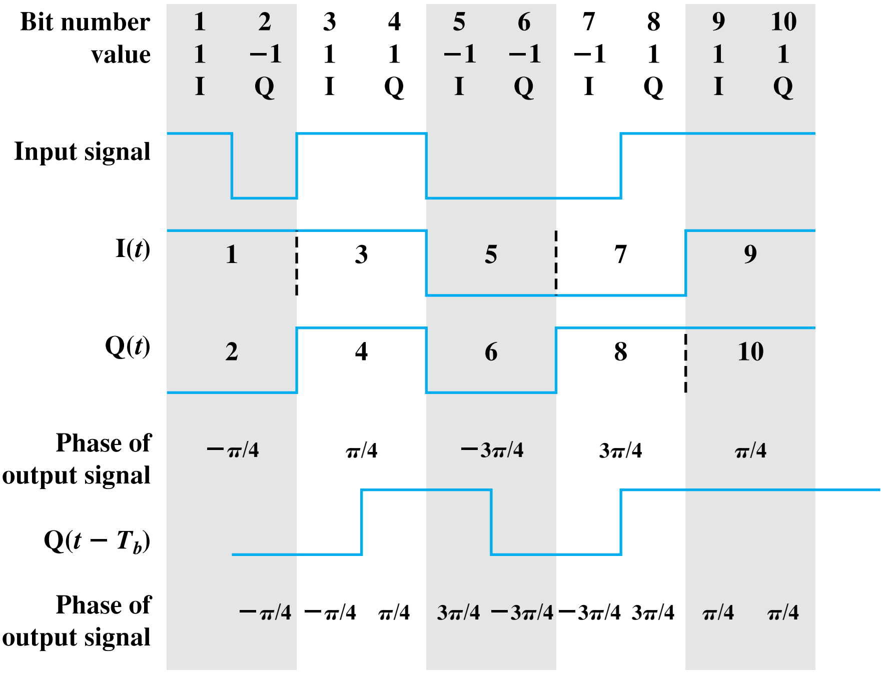


Figure 5.12 Example of QPSK and OQPSK Waveforms

From Figure 5.12, we can observe that only one of two bits in the pair can change sign at any time and thus the phase change in the combined signal never exceeds 90° ($\pi/2$). This can be an advantage because physical limitations on phase modulators make large phase shifts at high transition rates difficult to perform. OQPSK also provides superior performance when the transmission channel (including transmitter and receiver) has significant nonlinear components. The effect of nonlinearities is a spreading of the signal bandwidth, which may result in adjacent channel interference. It is easier to control this spreading if the phase changes are smaller, hence the advantage of OQPSK over QPSK.

Multilevel PSK The use of multiple levels can be extended beyond taking bits two at a time. It is possible to transmit bits three at a time using eight different phase angles. Further, each angle can have more than one amplitude. For example, a standard 9600 bps modem uses 12 phase angles, four of which have two amplitude values, for a total of 16 different signal elements.

This latter example points out very well the difference between the data rate R (in bps) and the modulation rate D (in baud) of a signal. Let us assume that this scheme is being employed with digital input in which each bit is represented by a constant voltage pulse, one level for binary one and one level for binary zero. The data rate is $R = 1/T_b$. However, the encoded signal contains $L = 4$ bits in each signal element using $M = 16$ different combinations of amplitude and phase. The modulation rate can be seen to be $R/4$, because each change of signal element communicates four bits. Thus the line signaling speed is 2400 baud, but the data rate is

9600 bps. This is the reason that higher bit rates can be achieved over voice-grade lines by employing more complex modulation schemes.

Performance

In looking at the performance of various digital-to-analog modulation schemes, the first parameter of interest is the bandwidth of the modulated signal. This depends on a variety of factors, including the definition of bandwidth used and the filtering technique used to create the bandpass signal. We will use some straightforward results from [COUC01].

The transmission bandwidth B_T for ASK is of the form

$$\textbf{ASK} \qquad B_T = (1 + r)R \tag{5.8}$$

where R is the bit rate and r is related to the technique by which the signal is filtered to establish a bandwidth for transmission; typically $0 < r < 1$. Thus the bandwidth is directly related to the bit rate. The preceding formula is also valid for PSK and, under certain assumptions, FSK.

With multilevel PSK (MPSK), significant improvements in bandwidth can be achieved. In general,

$$\textbf{MPSK} \qquad B_T = \left(\frac{1 + r}{L}\right)R = \left(\frac{1 + r}{\log_2 M}\right)R \tag{5.10}$$

where L is the number of bits encoded per signal element and M is the number of different signal elements.

For multilevel FSK (MFSK), we have

$$\textbf{MFSK} \qquad B_T = \left(\frac{(1 + r)M}{\log_2 M}\right)R \tag{5.11}$$

Table 5.5 shows the ratio of data rate, R, to transmission bandwidth for various schemes. This ratio is also referred to as the **bandwidth efficiency**. As the name suggests, this parameter measures the efficiency with which bandwidth can be used to transmit data. The advantage of multilevel signaling methods now becomes clear.

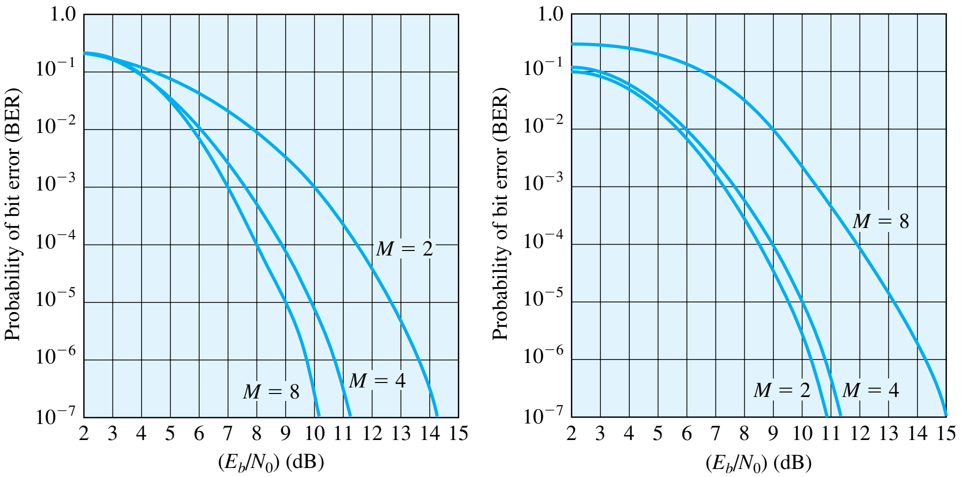
Of course, the preceding discussion refers to the spectrum of the input signal to a communications line. Nothing has yet been said of performance in the presence of noise. Figure 5.4 summarizes some results based on reasonable assumptions concerning the transmission system [COUC01]. Here bit error rate is plotted as a function of the ratio E_b/N_0 defined in Chapter 3. Of course, as that ratio increases, the bit error rate drops. Further, DPSK and BPSK are about 3 dB superior to ASK and BFSK.

Figure 5.13 shows the same information for various levels of M for MFSK and MPSK. There is an important difference. For MFSK, the error probability for a given value E_b/N_0 of decreases as M increases, while the opposite is true for MPSK. On the other hand, comparing Equations (5.10) and (5.11), the bandwidth efficiency of MFSK decreases as M increases, while the opposite is true of MPSK. Thus, in both

Table 5.5 Bandwidth Efficiency (R/B_T) for Various Digital-to-Analog Encoding Schemes

	$r = 0$	$r = 0.5$	$r = 1$
ASK	1.0	0.67	0.5
FSK	0.5	0.33	0.25
Multilevel FSK			
$M = 4, L = 2$	0.5	0.33	0.25
$M = 8, L = 3$	0.375	0.25	0.1875
$M = 16, L = 4$	0.25	0.167	0.125
$M = 32, L = 5$	0.156	0.104	0.078
PSK	1.0	0.67	0.5
Multilevel PSK			
$M = 4, L = 2$	2.00	1.33	1.00
$M = 8, L = 3$	3.00	2.00	1.50
$M = 16, L = 4$	4.00	2.67	2.00
$M = 32, L = 5$	5.00	3.33	2.50

cases, there is a tradeoff between bandwidth efficiency and error performance: An increase in bandwidth efficiency results in an increase in error probability. The fact that these tradeoffs move in opposite directions with respect to the number of levels *M* for MFSK and MPSK can be derived from the underlying equations. A discussion of the reasons for this difference is beyond the scope of this book. See [SKLA01] for a full treatment.



(a) Multilevel FSK (MFSK) **(b) Multilevel PSK (MPSK)**

Figure 5.13 Theoretical Bit Error Rate for Multilevel FSK and PSK

EXAMPLE 5.3 What is the bandwidth efficiency for FSK, ASK, PSK, and QPSK for a bit error rate of 10^{-7} on a channel with an SNR of 12 dB?

Using Equation (3.2), we have

$$\left(\frac{E_b}{N_0}\right)_{\text{dB}} = 12 \text{ dB} - \left(\frac{R}{B_T}\right)_{\text{dB}}$$

For FSK and ASK, from Figure 5.4,

$$\left(\frac{E_b}{N_0}\right)_{\text{dB}} = 14.2 \text{ dB}$$

$$\left(\frac{R}{B_T}\right)_{\text{dB}} = -2.2 \text{ dB}$$

$$\frac{R}{B_T} = 0.6$$

For PSK, from Figure 5.4,

$$\left(\frac{E_b}{N_0}\right)_{\text{dB}} = 11.2 \text{ dB}$$

$$\left(\frac{R}{B_T}\right)_{\text{dB}} = 0.8 \text{ dB}$$

$$\frac{R}{B_T} = 1.2$$

The result for QPSK must take into account that the baud rate $D = R/2$. Thus

$$\frac{R}{B_T} = 2.4$$

As the preceding example shows, ASK and FSK exhibit the same bandwidth efficiency, PSK is better, and even greater improvement can be achieved with multilevel signaling.

It is worthwhile to compare these bandwidth requirements with those for digital signaling. A good approximation is

$$B_T = 0.5(1 + r)D$$

where D is the modulation rate. For NRZ, $D = R$, and we have

$$\frac{R}{B_T} = \frac{2}{1 + r}$$

Thus digital signaling is in the same ballpark, in terms of bandwidth efficiency, as ASK, FSK, and PSK. A significant advantage for analog signaling is seen with multilevel techniques.

Quadrature Amplitude Modulation

Quadrature amplitude modulation (QAM) is a popular analog signaling technique that is used in the asymmetric digital subscriber line (ADSL), described in Chapter 8, and in some wireless standards. This modulation technique is a combination of ASK and PSK. QAM can also be considered a logical extension of QPSK. QAM takes advantage of the fact that it is possible to send two different signals simultaneously on the same carrier frequency, by using two copies of the carrier frequency, one shifted by 90° with respect to the other. For QAM, each carrier is ASK modulated. The two independent signals are simultaneously transmitted over the same medium. At the receiver, the two signals are demodulated and the results combined to produce the original binary input.

Figure 5.14 shows the QAM modulation scheme in general terms. The input is a stream of binary digits arriving at a rate of R bps. This stream is converted into two separate bit streams of $R/2$ bps each, by taking alternate bits for the two streams. In the diagram, the upper stream is ASK modulated on a carrier of frequency f_c by multiplying the bit stream by the carrier. Thus, a binary zero is represented by the absence of the carrier wave and a binary one is represented by the presence of the carrier wave at a constant amplitude. This same carrier wave is shifted by 90° and used for ASK modulation of the lower binary stream. The two modulated signals are then added together and transmitted. The transmitted signal can be expressed as follows:

$$\textbf{QAM} \qquad s(t) = d_1(t)\cos 2\pi f_c t + d_2(t)\sin 2\pi f_c t$$

If two-level ASK is used, then each of the two streams can be in one of two states and the combined stream can be in one of $4 = 2 \times 2$ states. This is essentially QPSK. If four-level ASK is used (i.e., four different amplitude levels), then the combined stream can be in one of $16 = 4 \times 4$ states. Systems using 64 and even 256 states have been implemented. The greater the number of states, the higher the data rate that is possible within a given bandwidth. Of course, as discussed previously, the greater the number of states, the higher the potential error rate due to noise and attenuation.

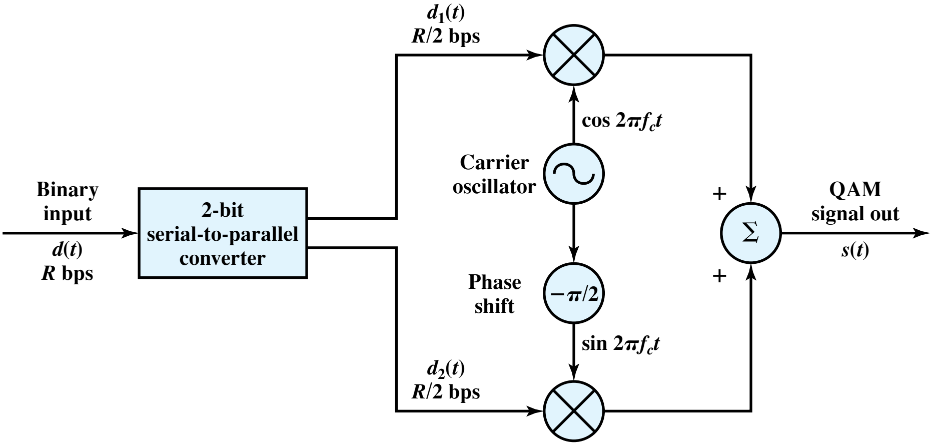


Figure 5.14 QAM Modulator

5.3 ANALOG DATA, DIGITAL SIGNALS

In this section we examine the process of transforming analog data into digital signals. Strictly speaking, it might be more correct to refer to this as a process of converting analog data into digital data; this process is known as digitization. Once analog data have been converted into digital data, a number of things can happen. The three most common are as follows:

1. The digital data can be transmitted using NRZ-L. In this case, we have in fact gone directly from analog data to a digital signal.
2. The digital data can be encoded as a digital signal using a code other than NRZ-L. Thus an extra step is required.
3. The digital data can be converted into an analog signal, using one of the modulation techniques discussed in Section 5.2.

This last, seemingly curious, procedure is illustrated in Figure 5.15, which shows voice data that are digitized and then converted to an analog ASK signal. This allows digital transmission in the sense defined in Chapter 3. The voice data, because they have been digitized, can be treated as digital data, even though transmission requirements (e.g., use of microwave) dictate that an analog signal be used.

The device used for converting analog data into digital form for transmission, and subsequently recovering the original analog data from the digital, is known as a **codec** (coder-decoder). In this section we examine the two principal techniques used in codecs, pulse code modulation and delta modulation. The section closes with a discussion of comparative performance.

Pulse Code Modulation

Pulse code modulation (PCM) is based on the sampling theorem:

> **SAMPLING THEOREM:** If a signal $f(t)$ is sampled at regular intervals of time and at a rate higher than twice the highest signal frequency, then the samples contain all the information of the original signal. The function $f(t)$ may be reconstructed from these samples by the use of a lowpass filter.

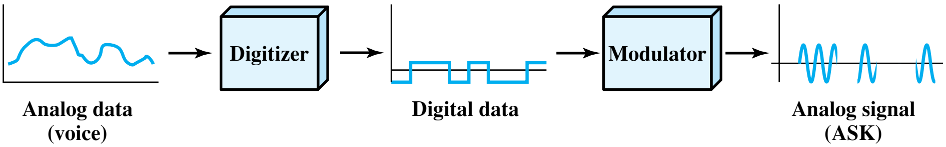
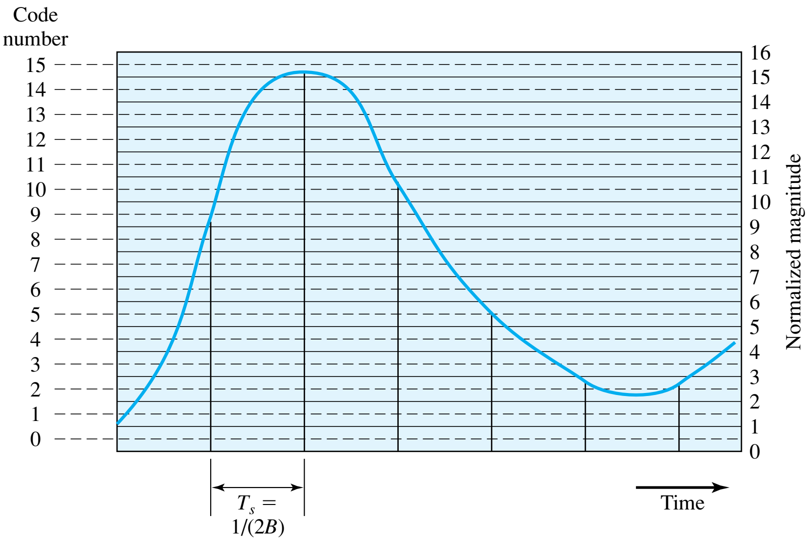


Figure 5.15 Digitizing Analog Data



PAM value	1.1	9.2	15.2	10.8	5.6	2.8	2.7
Quantized code number	1	9	15	10	5	2	2
PCM code	0001	1001	1111	1010	0101	0010	0010

Figure 5.16 Pulse Code Modulation Example

For the interested reader, a proof is provided in Appendix F. If voice data are limited to frequencies below 4000 Hz, a conservative procedure for intelligibility, 8000 samples per second would be sufficient to characterize the voice signal completely. Note, however, that these are analog samples, called **pulse amplitude modulation (PAM)** samples. To convert to digital, each of these analog samples must be assigned a binary code.

Figure 5.16 shows an example in which the original signal is assumed to be bandlimited with a bandwidth of B. PAM samples are taken at a rate of $2B$, or once every $T_s = 1/2B$ seconds. Each PAM sample is approximated by being *quantized* into one of 16 different levels. Each sample can then be represented by 4 bits. But because the quantized values are only approximations, it is impossible to recover the original signal exactly. By using an 8-bit sample, which allows 256 quantizing levels, the quality of the recovered voice signal is comparable with that achieved via analog transmission. Note that this implies that a data rate of 8000 samples per second × 8 bits per sample = 64 kbps is needed for a single voice signal.

Thus, PCM starts with a continuous-time, continuous-amplitude (analog) signal, from which a digital signal is produced (Figure 5.17). The digital signal consists of blocks of n bits, where each n-bit number is the amplitude of a PCM pulse. On reception, the process is reversed to reproduce the analog signal. Notice, however, that this process violates the terms of the sampling theorem. By quantizing the PAM pulse, the original signal is now only approximated and cannot be recovered exactly. This effect is known as **quantizing error** or **quantizing**

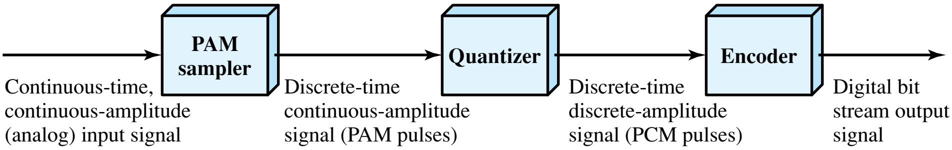

PAM sampler
Quantizer
Encoder
Continuous-time, continuous-amplitude (analog) input signal
Discrete-time continuous-amplitude signal (PAM pulses)
Discrete-time discrete-amplitude signal (PCM pulses)
Digital bit stream output signal

Figure 5.17 PCM Block Diagram

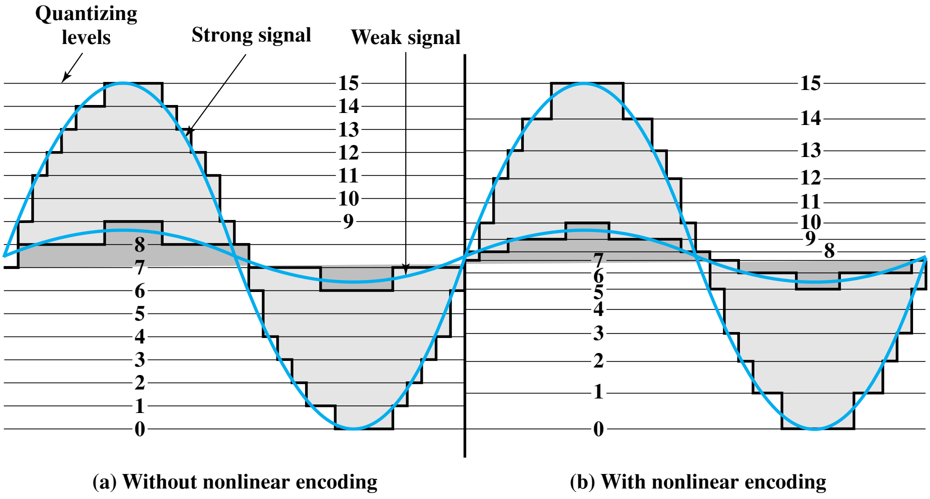
noise. The signal-to-noise ratio for quantizing noise can be expressed as [GIBS93]

$$\mathrm{SNR}_{\mathrm{dB}} = 20 \log 2^n + 1.76 \text{ dB} = 6.02n + 1.76 \text{ dB}$$

Thus each additional bit used for quantizing increases SNR by about 6 dB, which is a factor of 4.

Typically, the PCM scheme is refined using a technique known as nonlinear encoding, which means, in effect, that the quantization levels are not equally spaced. The problem with equal spacing is that the mean absolute error for each sample is the same, regardless of signal level. Consequently, lower amplitude values are relatively more distorted. By using a greater number of quantizing steps for signals of low amplitude, and a smaller number of quantizing steps for signals of large amplitude, a marked reduction in overall signal distortion is achieved (e.g., see Figure 5.18).

The same effect can be achieved by using uniform quantizing but companding (compressing-expanding) the input analog signal. Companding is a process that compresses the intensity range of a signal by imparting more gain to weak signals than to strong signals on input. At output, the reverse operation is performed. Figure 5.19 shows typical companding functions. Note that the effect on the input side is to compress the sample so that the higher values are reduced with respect



(a) Without nonlinear encoding

(b) With nonlinear encoding

Figure 5.18 Effect of Nonlinear Coding

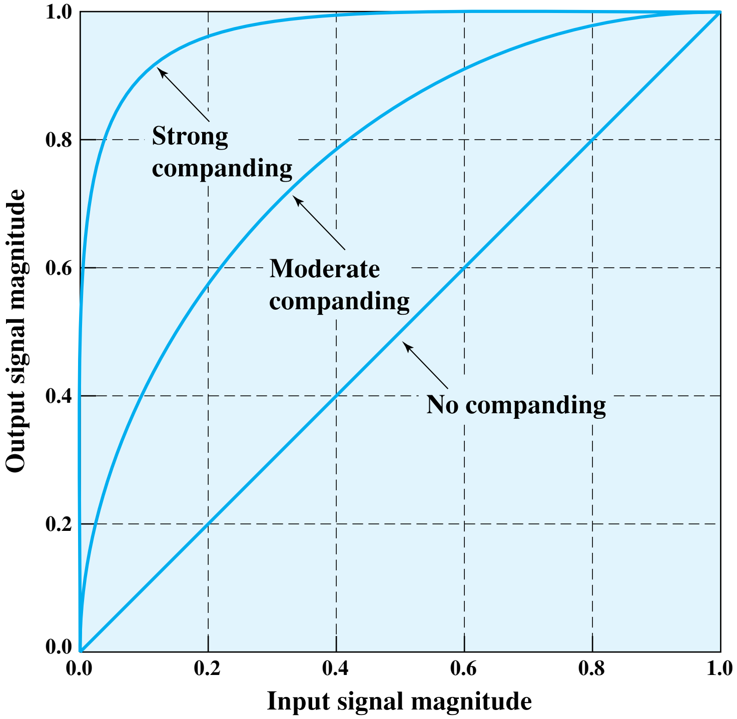


Figure 5.19 Typical Companding Functions

to the lower values. Thus, with a fixed number of quantizing levels, more levels are available for lower-level signals. On the output side, the compander expands the samples so the compressed values are restored to their original values.

Nonlinear encoding can significantly improve the PCM SNR ratio. For voice signals, improvements of 24 to 30 dB have been achieved.

Delta Modulation (DM)

A variety of techniques have been used to improve the performance of PCM or to reduce its complexity. One of the most popular alternatives to PCM is delta modulation (DM).

With delta modulation, an analog input is approximated by a staircase function that moves up or down by one quantization level (δ) at each sampling interval (T_s). An example is shown in Figure 5.20, where the staircase function is overlaid on the original analog waveform. The important characteristic of this staircase function is that its behavior is binary: At each sampling time, the function moves up or down a constant amount δ. Thus, the output of the delta modulation process can be represented as a single binary digit for each sample. In essence, a bit stream is produced by approximating the derivative of an analog signal rather than its amplitude: A 1 is generated if the staircase function is to go up during the next interval; a 0 is generated otherwise.

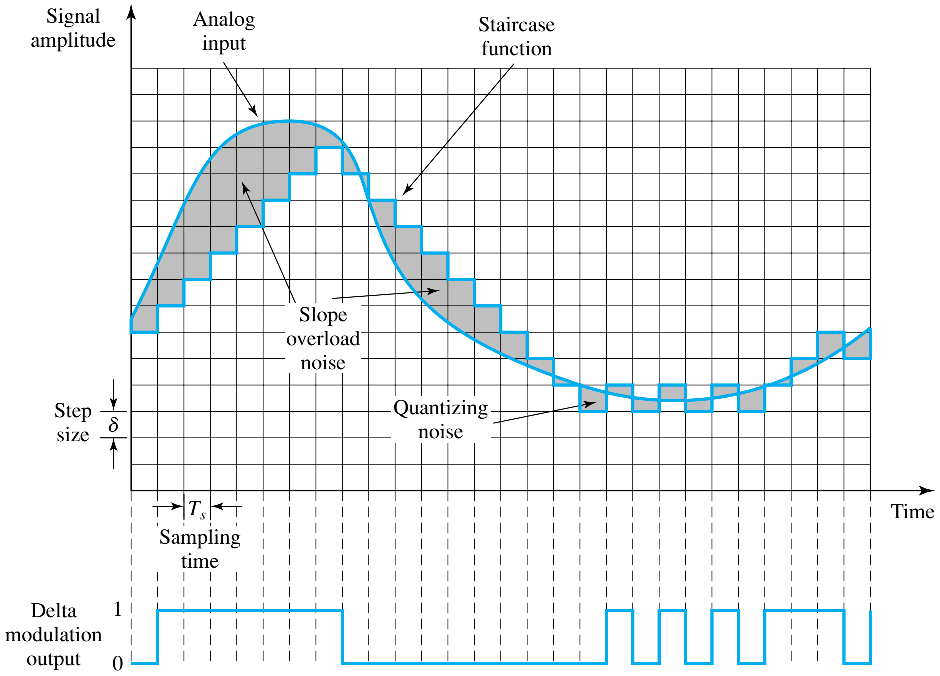


Figure 5.20 Example of Delta Modulation

The transition (up or down) that occurs at each sampling interval is chosen so that the staircase function tracks the original analog waveform as closely as possible. Figure 5.21 illustrates the logic of the process, which is essentially a feedback mechanism. For transmission, the following occurs: At each sampling time, the analog input is compared to the most recent value of the approximating staircase function. If the value of the sampled waveform exceeds that of the staircase function, a 1 is generated; otherwise, a 0 is generated. Thus, the staircase is always changed in the direction of the input signal. The output of the DM process is therefore a binary sequence that can be used at the receiver to reconstruct the staircase function. The staircase function can then be smoothed by some type of integration process or by passing it through a lowpass filter to produce an analog approximation of the analog input signal.

There are two important parameters in a DM scheme: the size of the step assigned to each binary digit, δ, and the sampling rate. As Figure 5.20 illustrates, δ must be chosen to produce a balance between two types of errors or noise. When the analog waveform is changing very slowly, there will be quantizing noise. This noise increases as δ is increased. On the other hand, when the analog waveform is changing more rapidly than the staircase can follow, there is slope overload noise. This noise increases as δ is decreased.

It should be clear that the accuracy of the scheme can be improved by increasing the sampling rate. However, this increases the data rate of the output signal.

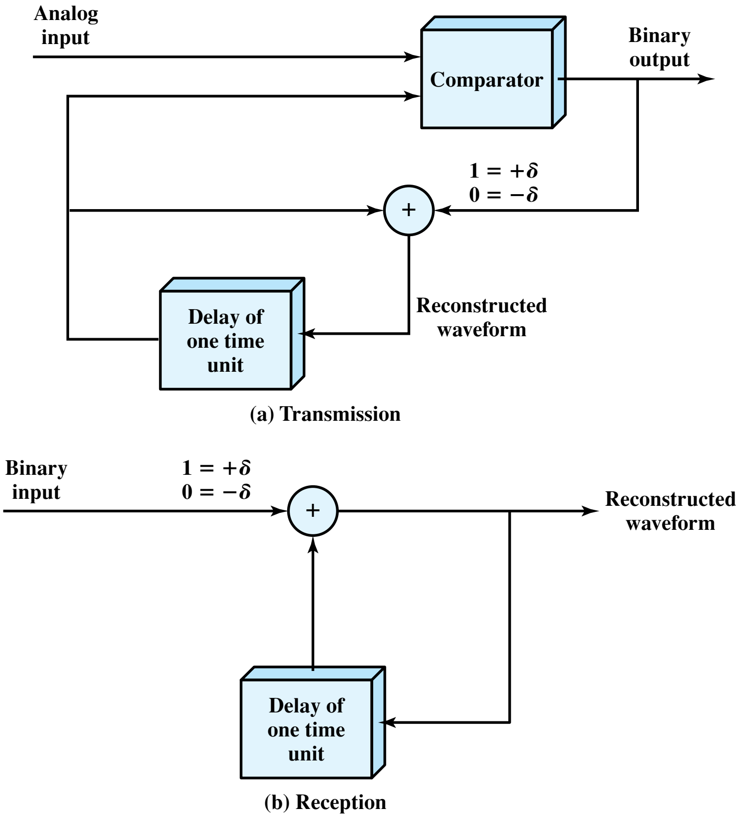


Figure 5.21 Delta Modulation

The principal advantage of DM over PCM is the simplicity of its implementation. In general, PCM exhibits better SNR characteristics at the same data rate.

Performance

Good voice reproduction via PCM can be achieved with 128 quantization levels, or 7-bit coding ($2^7 = 128$). A voice signal, conservatively, occupies a bandwidth of 4 kHz. Thus, according to the sampling theorem, samples should be taken at a rate of 8000 samples per second. This implies a data rate of $8000 \times 7 = 56$ kbps for the PCM-encoded digital data.

Consider what this means from the point of view of bandwidth requirement. An analog voice signal occupies 4 kHz. Using PCM this 4-kHz analog signal can be converted into a 56-kbps digital signal. But using the Nyquist criterion from Chapter 3, this digital signal could require on the order of 28 kHz of bandwidth. Even more severe differences are seen with higher bandwidth signals. For example, a common PCM scheme for color television uses 10-bit codes, which works out to 92 Mbps for a 4.6-MHz bandwidth signal. In spite of these numbers, digital techniques continue to grow in popularity for transmitting analog data. The principal reasons for this are as follows:

- Because repeaters are used instead of amplifiers, there is no cumulative noise.
- As we shall see, time division multiplexing (TDM) is used for digital signals instead of the frequency division multiplexing (FDM) used for analog signals. With TDM, there is no intermodulation noise, whereas we have seen that this is a concern for FDM.
- The conversion to digital signaling allows the use of the more efficient digital switching techniques.

Furthermore, techniques have been developed to provide more efficient codes. In the case of voice, a reasonable goal appears to be in the neighborhood of 4 kbps. With video, advantage can be taken of the fact that from frame to frame, most picture elements will not change. Interframe coding techniques should allow the video requirement to be reduced to about 15 Mbps, and for slowly changing scenes, such as found in a video teleconference, down to 64 kbps or less.

As a final point, we mention that in many instances, the use of a telecommunications system will result in both digital-to-analog and analog-to-digital processing. The overwhelming majority of local terminations into the telecommunications network is analog, and the network itself uses a mixture of analog and digital techniques. Thus digital data at a user's terminal may be converted to analog by a modem, subsequently digitized by a codec, and perhaps suffer repeated conversions before reaching its destination.

Thus, telecommunication facilities handle analog signals that represent both voice and digital data. The characteristics of the waveforms are quite different. Whereas voice signals tend to be skewed to the lower portion of the bandwidth (Figure 3.9), analog encoding of digital signals has a more uniform spectral content over the bandwidth and therefore contains more high-frequency components. Studies have shown that, because of the presence of these higher frequencies, PCM-related techniques are preferable to DM-related techniques for digitizing analog signals that represent digital data.

5.4 ANALOG DATA, ANALOG SIGNALS

Modulation has been defined as the process of combining an input signal $m(t)$ and a carrier at frequency f_c to produce a signal $s(t)$ whose bandwidth is (usually) centered on f_c. For digital data, the motivation for modulation should be clear: When only analog transmission facilities are available, modulation is required to convert the digital data to analog form. The motivation when the data are already analog is less clear. After all, voice signals are transmitted over telephone lines at their original spectrum (referred to as baseband transmission). There are two principal reasons for analog modulation of analog signals:

- A higher frequency may be needed for effective transmission. For unguided transmission, it is virtually impossible to transmit baseband signals; the required antennas would be many kilometers in diameter.
- Modulation permits frequency division multiplexing, an important technique explored in Chapter 8.

In this section we look at the principal techniques for modulation using analog data: amplitude modulation (AM), frequency modulation (FM), and phase modulation (PM). As before, the three basic characteristics of a signal are used for modulation.

Amplitude Modulation

Amplitude modulation (AM) is the simplest form of modulation and is depicted in Figure 5.22. Mathematically, the process can be expressed as

$$\textbf{AM} \qquad s(t) = [1 + n_a x(t)]\cos 2\pi f_c t \qquad \textbf{(5.12)}$$

where $\cos 2\pi f_c t$ is the carrier and $x(t)$ is the input signal (carrying data), both normalized to unity amplitude. The parameter n_a, known as the **modulation index**, is the ratio of the amplitude of the input signal to the carrier. Corresponding to our previous notation, the input signal is $m(t) = n_a x(t)$. The "1" in the Equation (5.12) is a dc component that prevents loss of information, as explained subsequently. This scheme is also known as double sideband transmitted carrier (DSBTC).

EXAMPLE 5.4 Derive an expression for $s(t)$ if $x(t)$ is the amplitude-modulating signal $\cos 2\pi f_m t$. We have

$$s(t) = [1 + n_a \cos 2\pi f_m t]\cos 2\pi f_c t$$

By trigonometric identity, this may be expanded to

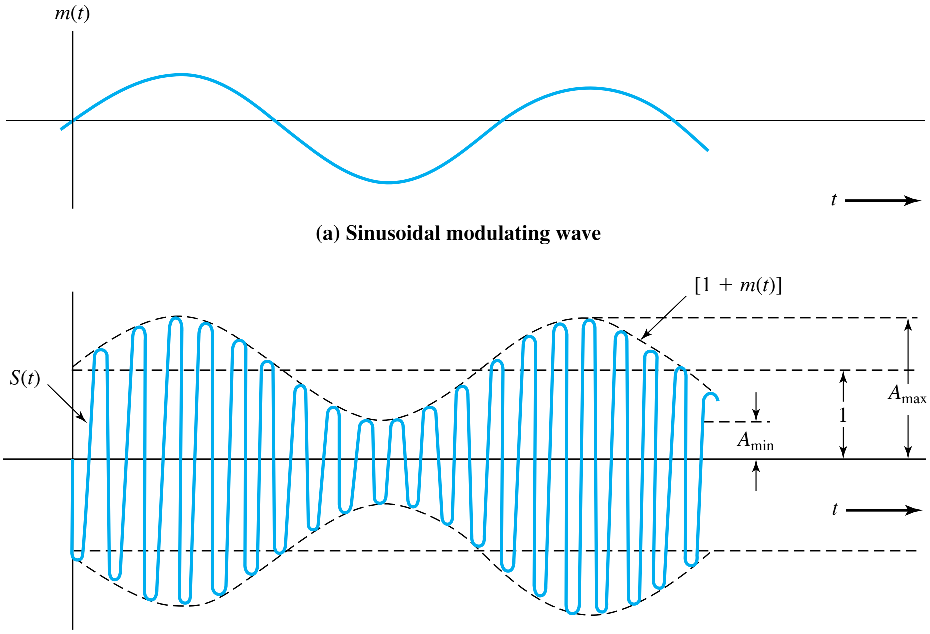
$$s(t) = \cos 2\pi f_c t + \frac{n_a}{2}\cos 2\pi(f_c - f_m)t + \frac{n_a}{2}\cos 2\pi(f_c + f_m)t$$

The resulting signal has a component at the original carrier frequency plus a pair of components each spaced f_m hertz from the carrier.

From Equation (5.12) and Figure 5.22, it can be seen that AM involves the multiplication of the input signal by the carrier. The envelope of the resulting signal is $[1 + n_a x(t)]$ and, as long as $n_a < 1$, the envelope is an exact reproduction of the original signal. If $n_a > 1$, the envelope will cross the time axis and information is lost.

It is instructive to look at the spectrum of the AM signal. An example is shown in Figure 5.23. The spectrum consists of the original carrier plus the spectrum of the input signal translated to f_c. The portion of the spectrum for $|f| > |f_c|$ is the *upper sideband,* and the portion of the spectrum for $|f| < |f_c|$ is *lower sideband.* Both the upper and lower sidebands are replicas of the original spectrum $M(f)$, with the lower sideband being frequency reversed. As an example, consider a voice signal with a bandwidth that extends from 300 to 3000 Hz being modulated on a 60-kHz carrier. The resulting signal contains an upper sideband of 60.3 to 63 kHz, a lower sideband of 57 to 59.7 kHz, and the 60-kHz carrier. An important relationship is

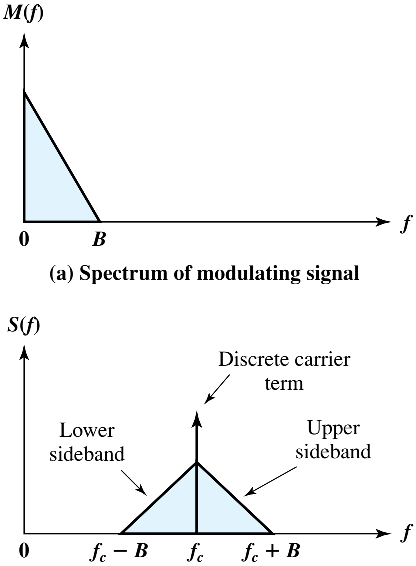
$$P_t = P_c\left(1 + \frac{n_a^2}{2}\right)$$



(a) Sinusoidal modulating wave

(b) Resulting AM signal

Figure 5.22 Amplitude Modulation



(a) Spectrum of modulating signal

(b) Spectrum of AM signal with carrier at f_c

Figure 5.23 Spectrum of an AM Signal

where P_t is the total transmitted power in $s(t)$ and P_c is the transmitted power in the carrier. We would like n_a as large as possible so that most of the signal power is used to carry information. However, n_a must remain below 1.

It should be clear that $s(t)$ contains unnecessary components, because each of the sidebands contains the complete spectrum of $m(t)$. A popular variant of AM, known as single sideband (SSB), takes advantage of this fact by sending only one of the sidebands, eliminating the other sideband and the carrier. The principal advantages of this approach are as follows:

- Only half the bandwidth is required, that is, $B_T = B$, where B is the bandwidth of the original signal. For DSBTC, $B_T = 2B$.
- Less power is required because no power is used to transmit the carrier or the other sideband. Another variant is double sideband suppressed carrier (DSBSC), which filters out the carrier frequency and sends both sidebands. This saves some power but uses as much bandwidth as DSBTC.

The disadvantage of suppressing the carrier is that the carrier can be used for synchronization purposes. For example, suppose that the original analog signal is an ASK waveform encoding digital data. The receiver needs to know the starting point of each bit time to interpret the data correctly. A constant carrier provides a clocking mechanism by which to time the arrival of bits. A compromise approach is vestigial sideband (VSB), which uses one sideband and a reduced-power carrier.

Angle Modulation

Frequency modulation (FM) and phase modulation (PM) are special cases of angle modulation. The modulated signal is expressed as

$$\textbf{Angle Modulation} \qquad s(t) = A_c \cos[2\pi f_c t + \phi(t)] \tag{5.13}$$

For phase modulation, the phase is proportional to the modulating signal:

$$\textbf{PM} \qquad \phi(t) = n_p m(t) \tag{5.14}$$

where n_p is the phase modulation index.

For frequency modulation, the derivative of the phase is proportional to the modulating signal:

$$\textbf{FM} \qquad \phi'(t) = n_f m(t) \tag{5.15}$$

where n_f is the frequency modulation index and $\phi'(t)$ is the derivative of $\phi(t)$.

For those who wish a more detailed mathematical explanation of the preceding, consider the following. The phase of $s(t)$ at any instant is just $2\pi f_c t + \phi(t)$. The instantaneous phase deviation from the carrier signal is $\phi(t)$. In PM, this instantaneous phase deviation is proportional to $m(t)$. Because frequency can be defined as the rate of change of phase of a signal, the instantaneous frequency of $s(t)$ is

$$2\pi f_i(t) = \frac{d}{dt}[2\pi f_c t + \phi(t)]$$

$$f_i(t) = f_c + \frac{1}{2\pi}\phi'(t)$$

and the instantaneous frequency deviation from the carrier frequency is $\phi'(t)$, which in FM is proportional to $m(t)$.

Figure 5.24 illustrates amplitude, phase, and frequency modulation by a sine wave. The shapes of the FM and PM signals are very similar. Indeed, it is impossible to tell them apart without knowledge of the modulation function.

Several observations about the FM process are in order. The peak deviation ΔF can be seen to be

$$\Delta F = \frac{1}{2\pi} n_f A_m \text{ Hz}$$

where A_m is the maximum value of $m(t)$. Thus an increase in the magnitude of $m(t)$ will increase ΔF, which, intuitively, should increase the transmitted bandwidth B_T. However, as should be apparent from Figure 5.24, this will not increase the average power level of the FM signal, which is $A_c^2/2$. This is distinctly different from AM, where the level of modulation affects the power in the AM signal but does not affect its bandwidth.

EXAMPLE 5.5 Derive an expression for $s(t)$ if $\phi(t)$ is the phase-modulating signal $n_p \cos 2\pi f_m t$. Assume that $A_c = 1$. This can be seen directly to be

$$s(t) = \cos[2\pi f_c t + n_p \cos 2\pi f_m t]$$

The instantaneous phase deviation from the carrier signal is $n_p \cos 2\pi f_m t$. The phase angle of the signal varies from its unmodulated value in a simple sinusoidal fashion, with the peak phase deviation equal to n_p.

The preceding expression can be expanded using Bessel's trigonometric identities:

$$s(t) = \sum_{n=-\infty}^{\infty} J_n(n_p) \cos\left(2\pi f_c t + 2\pi n f_m t + \frac{n\pi}{2}\right)$$

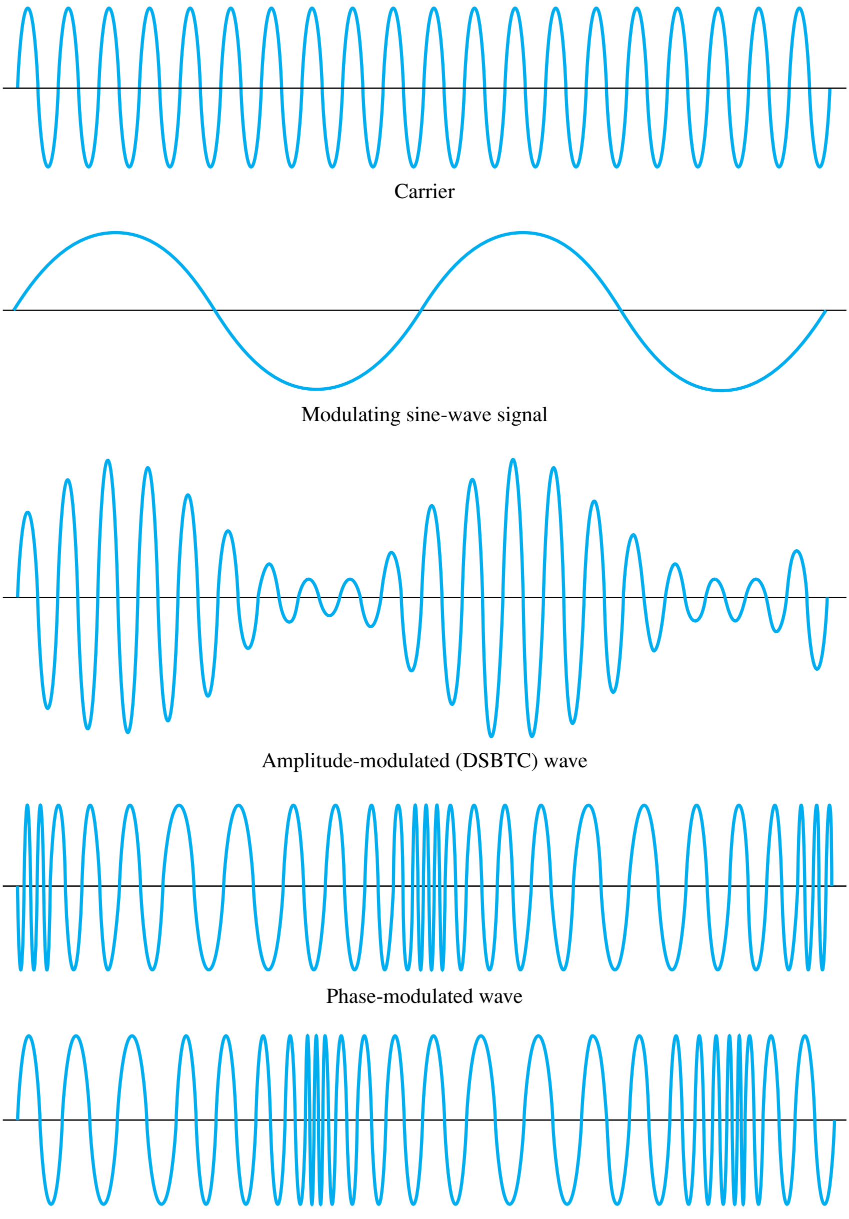
where $J_n(n_p)$ is the nth-order Bessel function of the first kind. Using the property

$$J_{-n}(x) = (-1)^n J_n(x)$$

this can be rewritten as

$$s(t) = J_0(n_p) \cos 2\pi f_c t + \sum_{n=1}^{\infty} J_n(n_p)\left[\cos\left(2\pi(f_c + n f_m)t + \frac{n\pi}{2}\right) + \cos\left(2\pi(f_c - n f_m)t + \frac{(n+2)\pi}{2}\right)\right]$$

The resulting signal has a component at the original carrier frequency plus a set of sidebands displaced from f_c by all possible multiples of f_m. For $n_p \ll 1$, the higher-order terms fall off rapidly.



Carrier

Modulating sine-wave signal

Amplitude-modulated (DSBTC) wave

Phase-modulated wave

Frequency-modulated wave

Figure 5.24 Amplitude, Phase, and Frequency Modulation of a Sine-Wave Carrier by a Sine-Wave Signal

EXAMPLE 5.6 Derive an expression for $s(t)$ if $\phi'(t)$ is the frequency modulating signal $-n_f \sin 2\pi f_m t$. The form of $\phi'(t)$ was chosen for convenience. We have

$$\phi(t) = -\int n_f \sin 2\pi f_m t \, dt = \frac{n_f}{2\pi f_m} \cos 2\pi f_m t$$

Thus

$$\begin{aligned} s(t) &= \cos\left[2\pi f_c t + \frac{n_f}{2\pi f_m} \cos 2\pi f_m t\right] \\ &= \cos\left[2\pi f_c t + \frac{\Delta F}{f_m} \cos 2\pi f_m t\right] \end{aligned}$$

The instantaneous frequency deviation from the carrier signal is $-n_f \sin 2\pi f_m t$. The frequency of the signal varies from its unmodulated value in a simple sinusoidal fashion, with the peak frequency deviation equal to n_f radians/second.

The equation for the FM signal has the identical form as for the PM signal, with $\Delta F/f_m$ substituted for n_p. Thus the Bessel expansion is the same.

As with AM, both FM and PM result in a signal whose bandwidth is centered at f_c. However, we can now see that the magnitude of that bandwidth is very different. Amplitude modulation is a linear process and produces frequencies that are the sum and difference of the carrier signal and the components of the modulating signal. Hence, for AM,

$$B_T = 2B$$

However, angle modulation includes a term of the form $\cos(\phi(t))$, which is nonlinear and will produce a wide range of frequencies. In essence, for a modulating sinusoid of frequency f_m, $s(t)$ will contain components at $f_c + f_m$, $f_c + 2f_m$, and so on. In the most general case, infinite bandwidth is required to transmit an FM or PM signal. As a practical matter, a very good rule of thumb, known as Carson's rule [COUC01], is

$$B_T = 2(\beta + 1)B$$

where

$$\beta = \begin{cases} n_p A_m & \text{for PM} \\ \dfrac{\Delta F}{B} = \dfrac{n_f A_m}{2\pi B} & \text{for FM} \end{cases}$$

We can rewrite the formula for FM as

$$B_T = 2\Delta F + 2B \tag{5.16}$$

Thus both FM and PM require greater bandwidth than AM.

5.5 RECOMMENDED READING

It is difficult, for some reason, to find solid treatments of digital-to-digital encoding schemes. Useful accounts include [SKLA01] and [BERG96].

There are many good references on analog modulation schemes for digital data. Good choices are [COUC01], [XION00], and [PROA05]; these three also provide comprehensive treatment of digital and analog modulation schemes for analog data.

An instructive treatment of the concepts of bit rate, baud, and bandwidth is [FREE98]. A recommended tutorial that expands on the concepts treated in the past few chapters relating to bandwidth efficiency and encoding schemes is [SKLA93].

BERG96 Bergmans, J. *Digital Baseband Transmission and Recording.* Boston: Kluwer, 1996.

COUC01 Couch, L. *Digital and Analog Communication Systems.* Upper Saddle River, NJ: Prentice Hall, 2001.

FREE98 Freeman, R. "Bits, Symbols, Baud, and Bandwidth." *IEEE Communications Magazine*, April 1998.

PROA05 Proakis, J. *Fundamentals of Communication Systems.* Upper Saddle River, NJ: Prentice Hall, 2005.

SKLA93 Sklar, B. "Defining, Designing, and Evaluating Digital Communication Systems." *IEEE Communications Magazine*, November 1993.

SKLA01 Sklar, B. *Digital Communications: Fundamentals and Applications.* Englewood Cliffs, NJ: Prentice Hall, 2001.

XION00 Xiong, F. *Digital Modulation Techniques.* Boston: Artech House, 2000.

5.6 KEY TERMS, REVIEW QUESTIONS, AND PROBLEMS

Key Terms

alternate mark inversion (AMI)
amplitude modulation (AM)
amplitude shift keying (ASK)
angle modulation
bandwidth efficiency
baseband signal
biphase
bipolar-AMI
bipolar with 8-zeros substitution (B8ZS)
bit error rate (BER)
carrier frequency
delta modulation (DM)
differential encoding
differential Manchester
differential PSK (DPSK)
frequency modulation (FM)
frequency shift keying (FSK)
high-density bipolar-3 zeros (HDB3)
Manchester
modulation
modulation rate
multilevel binary
nonreturn to zero (NRZ)
nonreturn to zero, inverted (NRZI)
nonreturn to zero-level (NRZ-L)
phase modulation (PM)
phase shift keying (PSK)
polar
pseudoternary
pulse amplitude modulation (PAM)
pulse code modulation (PCM)
quadrature amplitude modulation (QAM)
quadrature PSK (QPSK)
scrambling
unipolar

Review Questions

5.1. List and briefly define important factors that can be used in evaluating or comparing the various digital-to-digital encoding techniques.

5.2. What is differential encoding?

5.3. Explain the difference between NRZ-L and NRZI.

5.4. Describe two multilevel binary digital-to-digital encoding techniques.

5.5. Define biphase encoding and describe two biphase encoding techniques.

5.6. Explain the function of scrambling in the context of digital-to-digital encoding techniques.

5.7. What function does a modem perform?

5.8. How are binary values represented in amplitude shift keying, and what is the limitation of this approach?

5.9. What is the difference between QPSK and offset QPSK?

5.10. What is QAM?

5.11. What does the sampling theorem tell us concerning the rate of sampling required for an analog signal?

5.12. What are the differences among angle modulation, PM, and FM?

Problems

5.1 Which of the signals of Table 5.2 use differential encoding?

5.2 Develop algorithms for generating each of the codes of Table 5.2 from NRZ-L.

5.3 A modified NRZ code known as enhanced-NRZ (E-NRZ) is sometimes used for high-density magnetic tape recording. E-NRZ encoding entails separating the NRZ-L data stream into 7-bit words; inverting bits 2, 3, 6, and 7; and adding one parity bit to each word. The parity bit is chosen to make the total number of 1s in the 8-bit word an odd count. What are the advantages of E-NRZ over NRZ-L? Any disadvantages?

5.4 Develop a state diagram (finite state machine) representation of pseudoternary coding.

5.5 Consider the following signal encoding technique. Binary data are presented as input, a_m, for $m = 1, 2, 3, \ldots$ Two levels of processing occur. First, a new set of binary numbers are produced:

$$b_0 = 0$$
$$b_m = (a_m + b_{m-1}) \bmod 2$$

These are then encoded as

$$c_m = b_m - b_{m-1}$$

On reception, the original data are recovered by

$$a_m = c_m \bmod 2$$

a. Verify that the received values of a_m equal the transmitted values of a_m.

b. What sort of encoding is this?

5.6 For the bit stream 01001110, sketch the waveforms for each of the codes of Table 5.2. Assume that the signal level for the preceding bit for NRZI was high; the most recent preceding 1 bit (AMI) has a negative voltage; and the most recent preceding 0 bit (pseudoternary) has a negative voltage.

5.7 The waveform of Figure 5.25 belongs to a Manchester encoded binary data stream. Determine the beginning and end of bit periods (i.e., extract clock information) and give the data sequence.

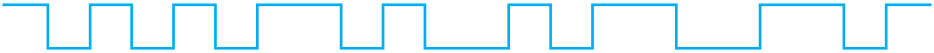


Figure 5.25 A Manchester Stream

5.8 Consider a stream of binary data consisting of a long sequence of 1s followed by a zero followed by a long string of 1s, with the same assumptions as Problem 5.6. Draw the waveform for this sequence using

a. NRZ-L

b. Bipolar-AMI

c. Pseudoternary

5.9 The bipolar-AMI waveform representing the binary sequence 0100101011 is transmitted over a noisy channel. The received waveform is shown in Figure 5.26; it contains a single error. Locate the position of this error and explain your answer.

5.10 One positive side effect of bipolar encoding is that a bipolar violation (two consecutive + pulses or two consecutive − pulses separated by any number of zeros) indicates to the receiver that an error has occurred in transmission. Unfortunately, upon the receipt of such a violation, the receiver does not know which bit is in error (only that an error has occurred). For the received bipolar sequence

$$+ - 0 + - 0 - +$$

which has one bipolar violation, construct two scenarios (each of which involves a different transmitted bit stream with one transmitted bit being converted via an error) that will produce this same received bit pattern.

5.11 Given the bit pattern 01100, encode this data using ASK, BFSK, and BPSK.

5.12 A sine wave is to be used for two different signaling schemes: (a) PSK; (b) QPSK. The duration of a signal element is 10^{-5} s. If the received signal is of the following form:

$$s(t) = 0.005 \sin(2\pi\, 10^6 t + \theta) \text{ volts}$$

and if the measured noise power at the receiver is 2.5×10^{-8} watts, determine the E_b/N_0 (in dB) for each case.

5.13 Derive an expression for baud rate D as a function of bit rate R for QPSK using the digital encoding techniques of Table 5.2.

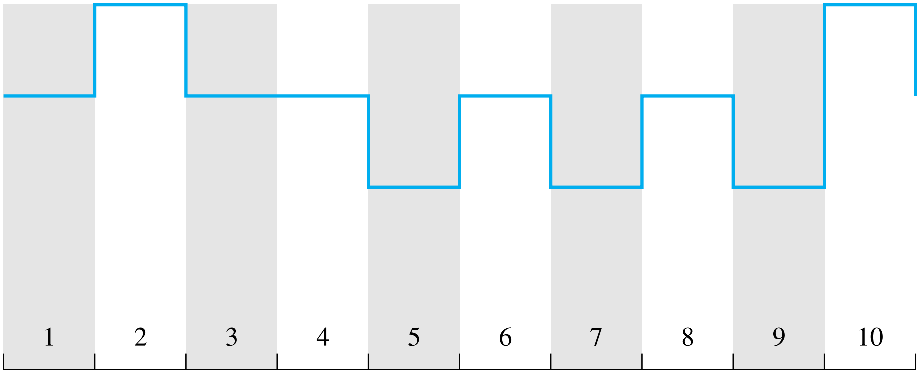


Figure 5.26 A Received Bipolar-AMI Waveform

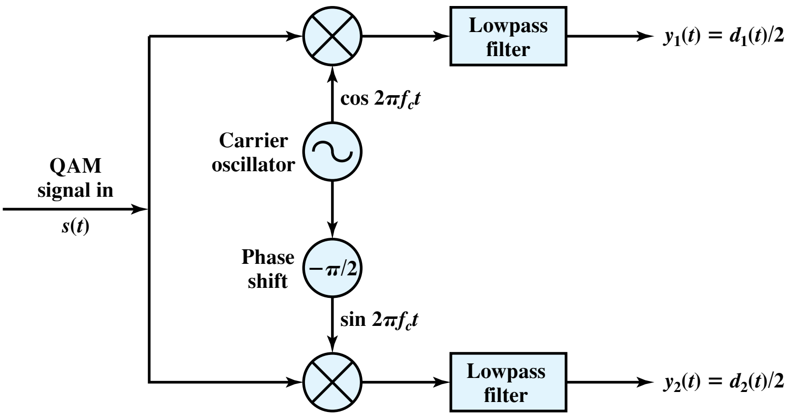


Figure 5.27 QAM Demodulator

5.14 What SNR ratio is required to achieve a bandwidth efficiency of 1.0 for ASK, FSK, PSK, and QPSK? Assume that the required bit error rate is 10^{-6}.

5.15 An NRZ-L signal is passed through a filter with $r = 0.5$ and then modulated onto a carrier. The data rate is 2400 bps. Evaluate the bandwidth for ASK and FSK. For FSK assume that the two frequencies used are 50 kHz and 55 kHz.

5.16 Assume that a telephone line channel is equalized to allow bandpass data transmission over a frequency range of 600 to 3000 Hz. The available bandwidth is 2400 Hz. For $r = 1$, evaluate the required bandwidth for 2400 bps QPSK and 4800-bps, eight-level multilevel signaling. Is the bandwidth adequate?

5.17 Figure 5.27 shows the QAM demodulator corresponding to the QAM modulator of Figure 5.14. Show that this arrangement does recover the two signals $d_1(t)$ and $d_2(t)$, which can be combined to recover the original input.

5.18 Why should PCM be preferable to DM for encoding analog signals that represent digital data?

5.19 Are the modem and the codec functional inverses (i.e., could an inverted modem function as a codec, or vice versa)?

5.20 A signal is quantized using 10-bit PCM. Find the signal-to-quantization noise ratio.

5.21 Consider an audio signal with spectral components in the range 300 to 3000 Hz. Assume that a sampling rate of 7000 samples per second will be used to generate a PCM signal.

a. For SNR = 30 dB, what is the number of uniform quantization levels needed?
b. What data rate is required?

5.22 Find the step size δ required to prevent slope overload noise as a function of the frequency of the highest-frequency component of the signal. Assume that all components have amplitude A.

5.23 A PCM encoder accepts a signal with a full-scale voltage of 10 V and generates 8-bit codes using uniform quantization. The maximum normalized quantized voltage is $1 - 2^{-8}$. Determine (a) normalized step size, (b) actual step size in volts, (c) actual maximum quantized level in volts, (d) normalized resolution, (e) actual resolution, and (f) percentage resolution.

5.24 The analog waveform shown in Figure 5.28 is to be delta modulated. The sampling period and the step size are indicated by the grid on the figure. The first DM output and the staircase function for this period are also shown. Show the rest of the staircase function and give the DM output. Indicate regions where slope overload distortion exists.

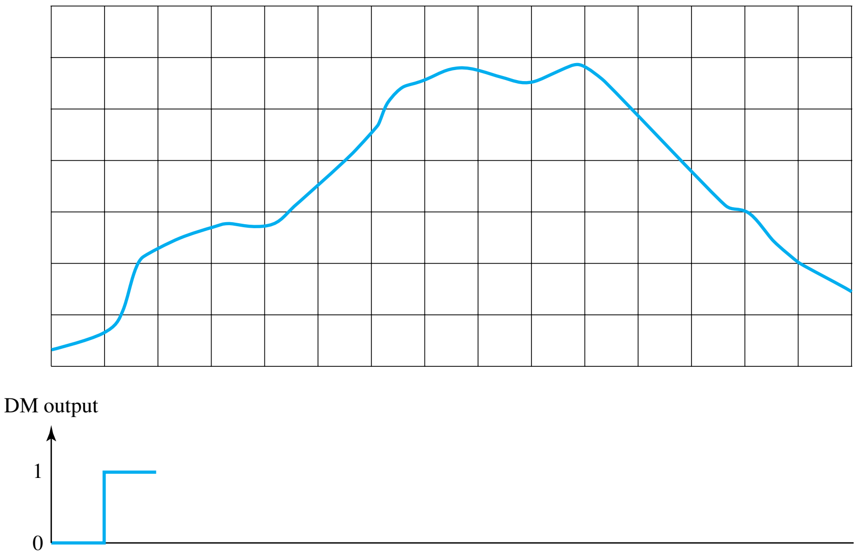


Figure 5.28 Delta Modulation Example

5.25 Consider the angle-modulated signal

$$s(t) = 10 \cos[(10^8)\pi t + 5 \sin 2\pi(10^3)t]$$

Find the maximum phase deviation and the maximum frequency deviation.

5.26 Consider the angle-modulated signal

$$s(t) = 10 \cos[2\pi(10^6)t + 0.1 \sin(10^3)\pi t]$$

a. Express $s(t)$ as a PM signal with $n_p = 10$.

b. Express $s(t)$ as an FM signal with $n_f = 10\pi$.

5.27 Let $m_1(t)$ and $m_2(t)$ be message signals and let $s_1(t)$ and $s_2(t)$ be the corresponding modulated signals using a carrier frequency of f_c.

a. Show that if simple AM modulation is used, then $m_1(t) + m_2(t)$ produces a modulated signal equal that is a linear combination of $s_1(t)$ and $s_2(t)$. This is why AM is sometimes referred to as linear modulation.

b. Show that if simple PM modulation is used, then $m_1(t) + m_2(t)$ produces a modulated signal that is not a linear combination of $s_1(t)$ and $s_2(t)$. This is why angle modulation is sometimes referred to as nonlinear modulation.