PART TWO: DATA COMMUNICATIONS

CHAPTER 3

DATA TRANSMISSION

LEARNING OBJECTIVES

After studying this chapter, you should be able to:

- Distinguish between digital and analog information sources.
- Explain the various ways in which audio, data, image, and video can be represented by electromagnetic signals.
- Discuss the characteristics of analog and digital waveforms.
- Discuss the various transmission impairments that affect signal quality and information transfer over communication media.
- Identify the factors that affect channel capacity.

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. 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*, 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 1.5, 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 varies in a smooth, or **continuous**, 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, in a **discrete** fashion.[2] Figure 3.1 shows an example of each kind of signal. The analog signal might represent speech, and the digital signal might represent binary 1s and 0s.

---

[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.

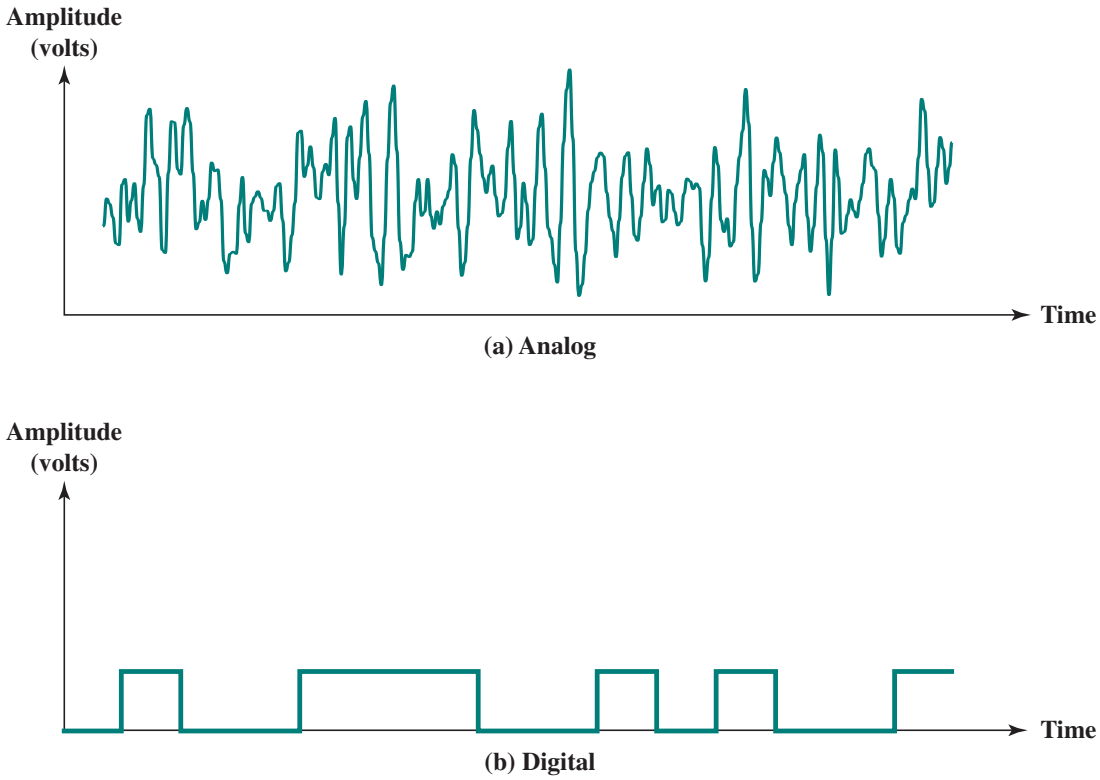


Figure 3.1 Analog and Digital Waveforms

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 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

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

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

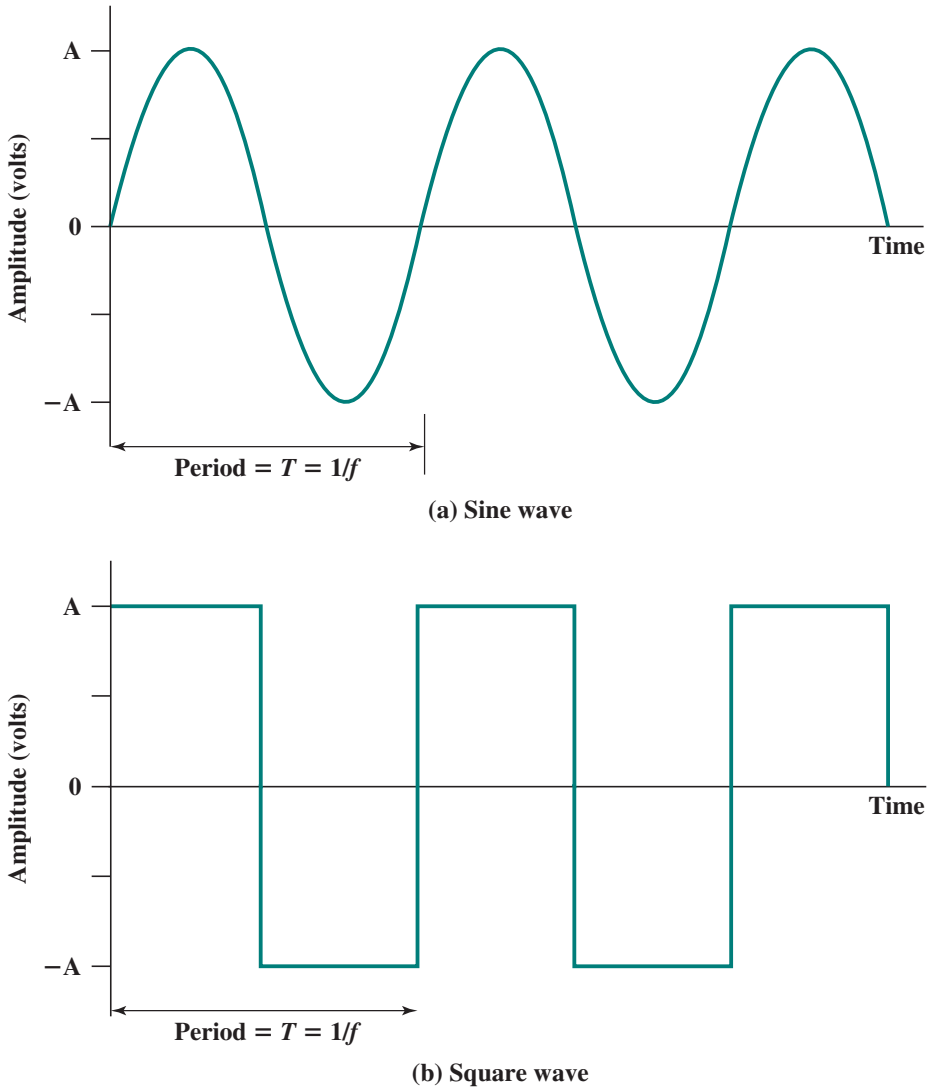


Figure 3.2 Examples of Periodic Signals

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).

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.[3]

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

[3]An electromagnetic signal attenuates as it propagates, as a function of distance from the source of the signal. This effect is ignored in Figure 3.3.

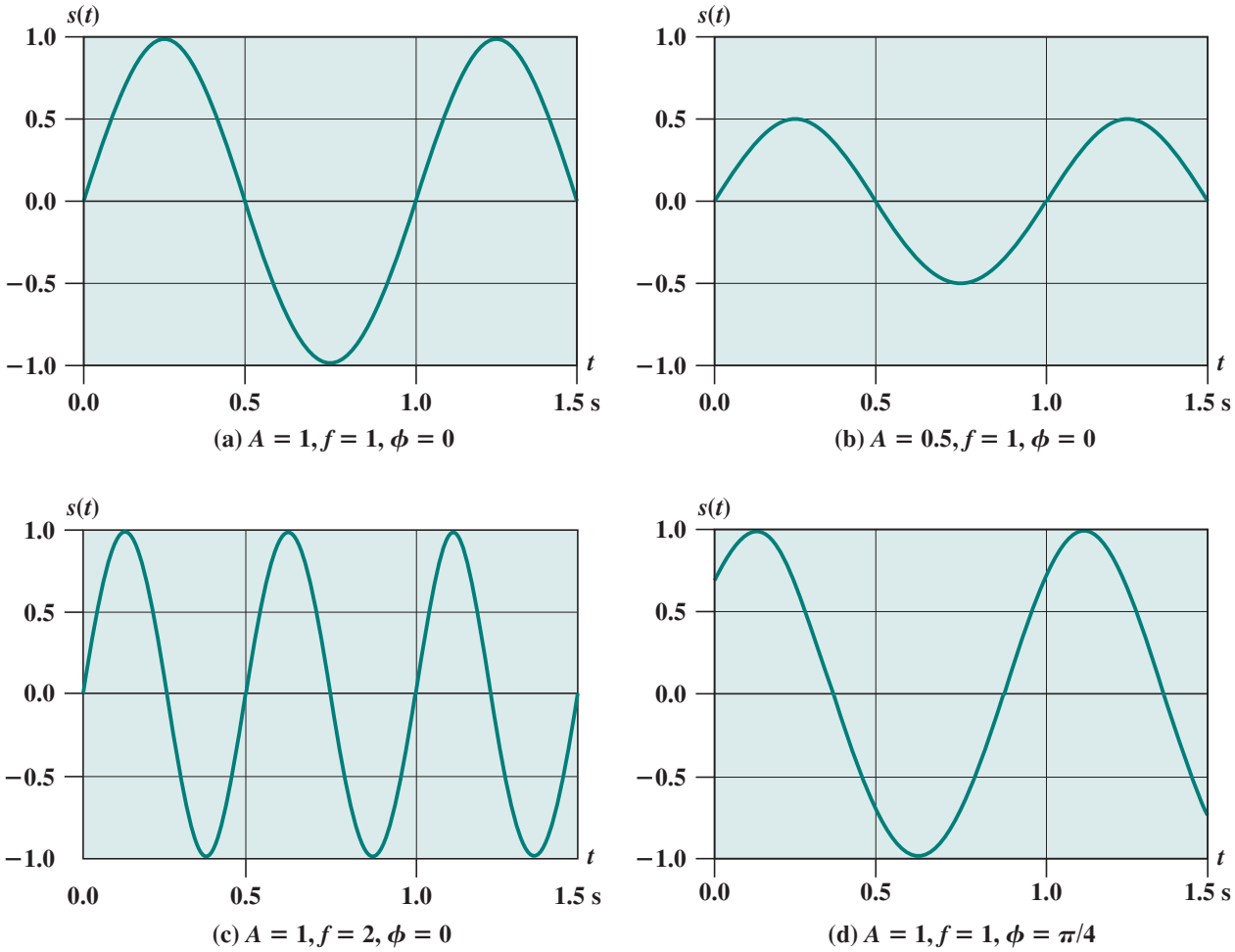


Figure 3.3 $s(t) = A \sin(2\pi ft + \phi)$

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.

EXAMPLE 3.1 In the United States, ordinary household current is typically supplied at a frequency of 60 Hz with a peak voltage of about 170 V. Thus the power line voltage can be expressed as

$$170 \sin(2\pi \times 60 \times t)$$

The period of this current is $1/60 = 0.0167$ s $= 16.7$ ms. A typical velocity of propagation is about 0.9 c, so the wavelength of the current is $\lambda = vT = 0.9 \times 3 \times 10^8 \times 0.0167 = 4.5 \times 10^6$ m $= 4500$ km.

Household voltage is normally stated as being 120 V. This is what is known as the root mean square (square the voltage to make everything positive, find the average, take the square root) value. For a sine wave, the value is calculated as $\sqrt{(A^2 - 0)/2} = 0.707A$. In this case, $0.707 \times 170 = 120$.

FREQUENCY DOMAIN CONCEPTS In practice, an electromagnetic signal will be made up of many frequencies. For example, the signal

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

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.[4] There are two interesting points that can be made about this figure:

- 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**. Each multiple of the fundamental frequency is referred to as a **harmonic frequency** of the signal.
- 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.[5] 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 $3f - f = 2f$ 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**.

---

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

[5]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.

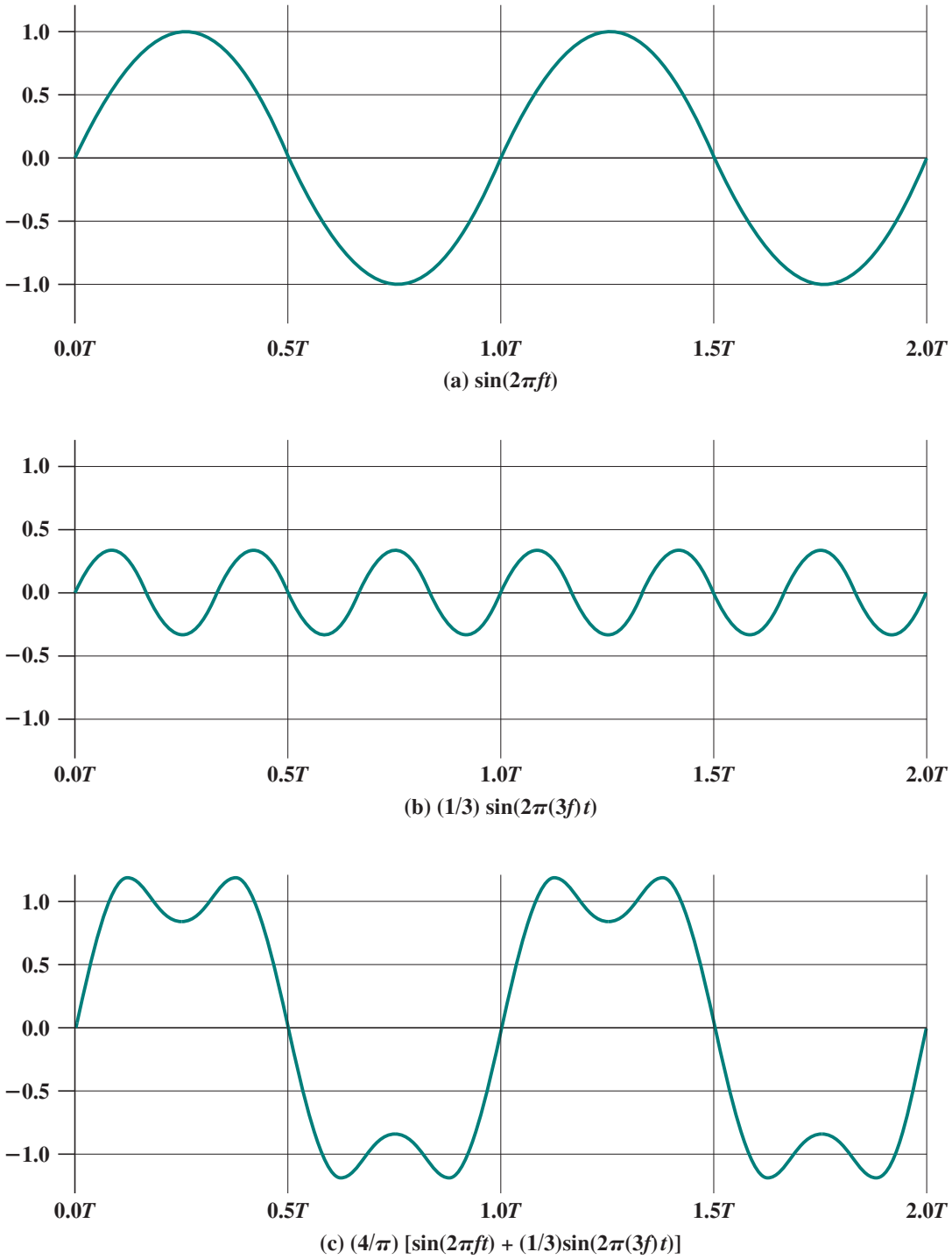
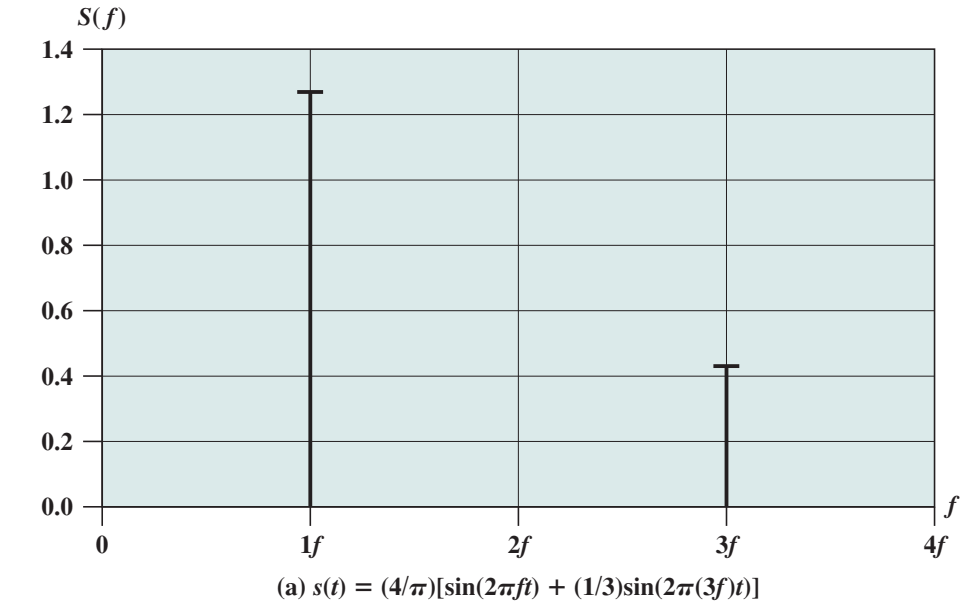
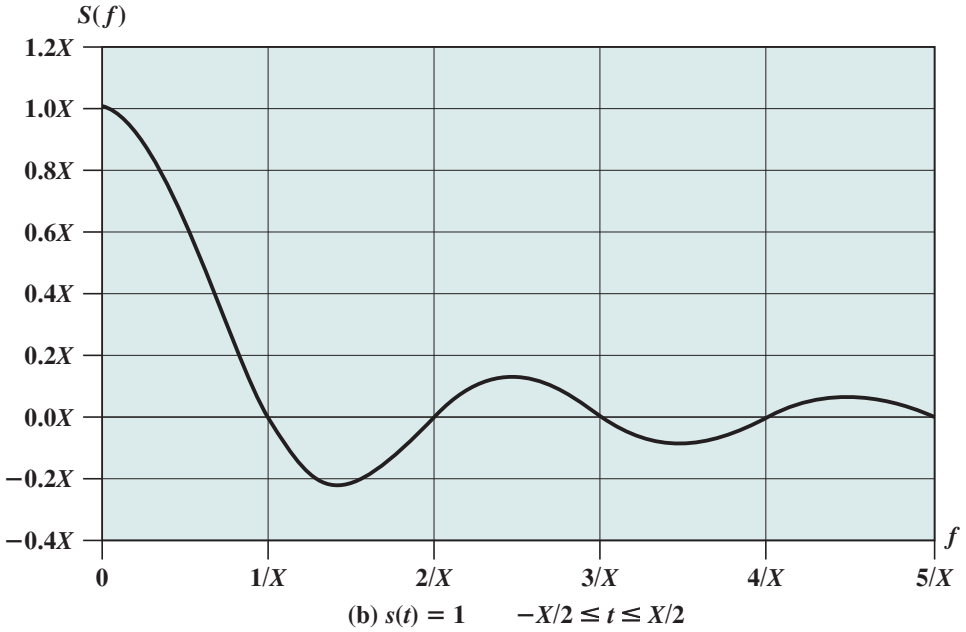


Figure 3.4 Addition of Frequency Components ($T = 1/f$)

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.



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



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

Figure 3.5 Frequency Domain Representations

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 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

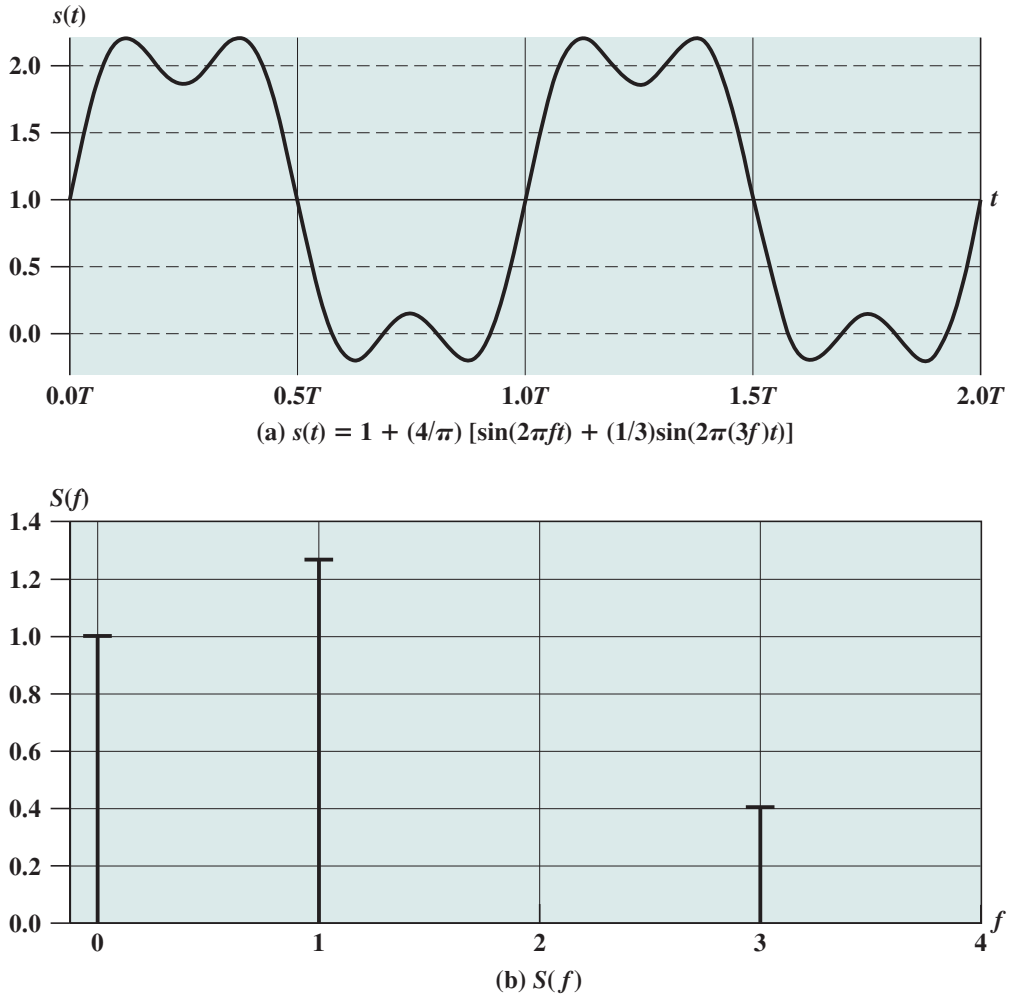


Figure 3.6 Signal with dc Component

represent binary 1. Then the waveform represents the repetitive 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 Figure 3.4. By adding together sine waves at frequencies f and $3f$, we get a waveform that begins to resemble the original square wave. Let us continue this process by adding a sine wave of frequency $5f$, as shown in Figure 3.7a, and then adding a sine wave of frequency $7f$, 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:

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

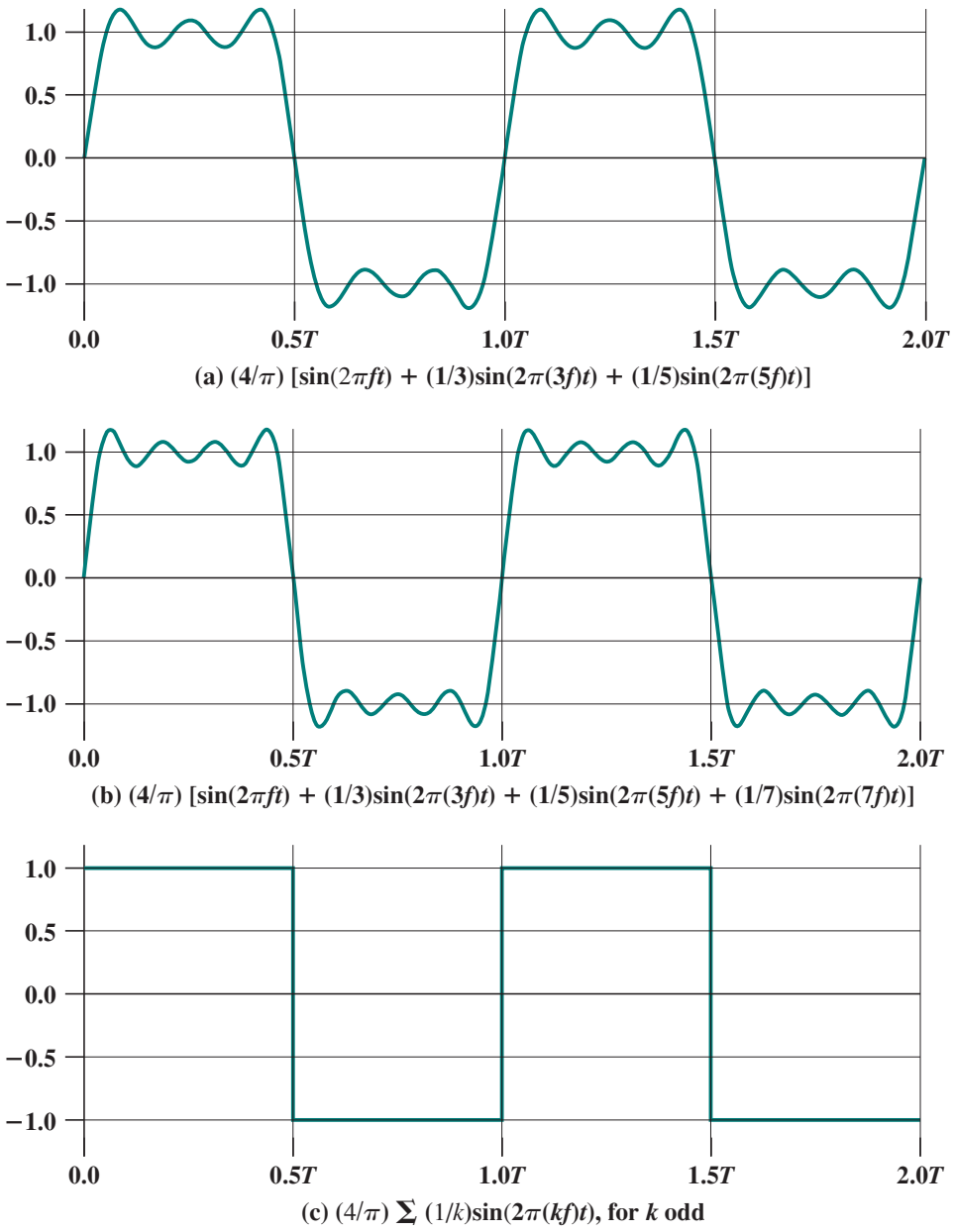


Figure 3.7 Frequency Components of Square Wave ($T = 1/f$)

Thus, this waveform has an infinite number of frequency components and hence an infinite bandwidth. However, the peak amplitude of the kth 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 see, the shape of the resulting waveform is reasonably close to that of the original square wave.

EXAMPLE 3.2 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, which consists of the fundamental frequency and two harmonics. 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

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

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, 1 bit occurs every $0.5\,\mu$s, 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, 1 bit occurs every $0.25\,\mu$s 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 1 bit occurs every $0.25\,\mu$s 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 example. 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

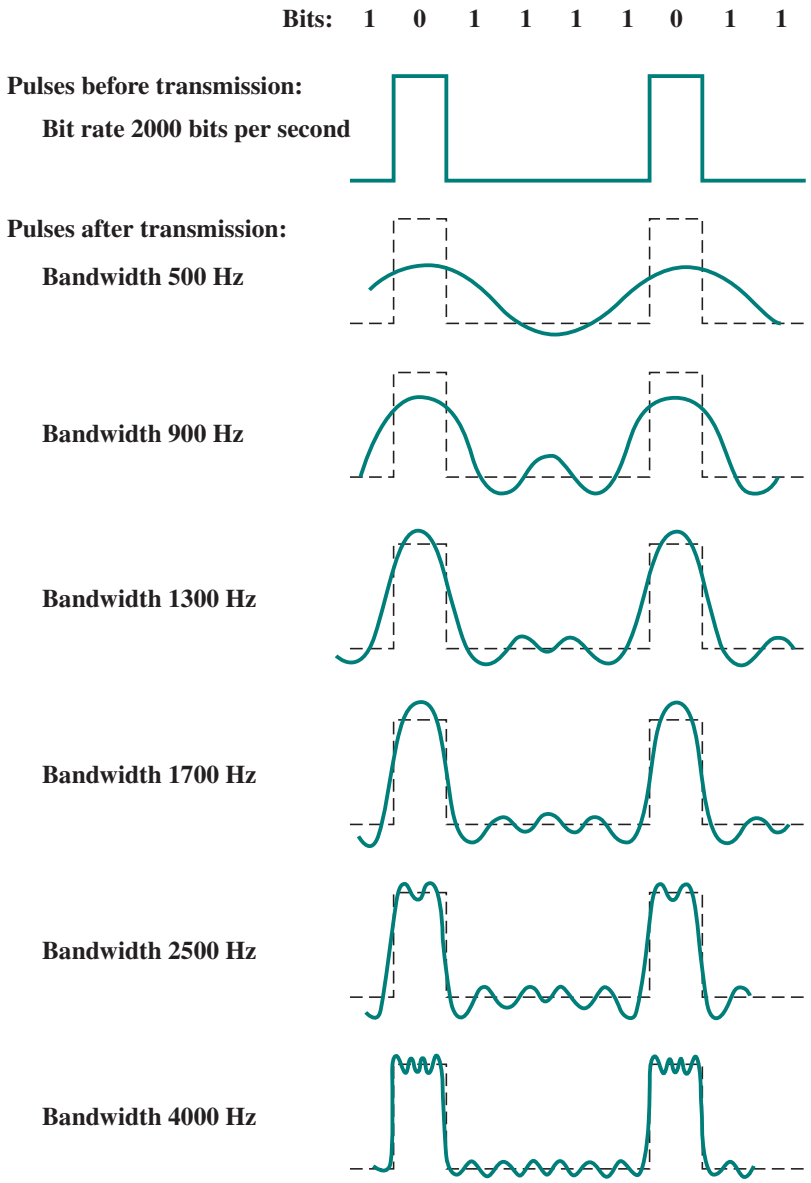


Figure 3.8 Effect of Bandwidth on a Digital Signal

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 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 the bandwidth required for transmission.

Another way of stating this is that the greater the bandwidth of a transmission system, the higher 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.[6] 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;[7] that is, the power produced by the loudest shout may be as much as 300 times greater than the least whisper.

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

[6]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 http://www.computersciencestudent.com

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

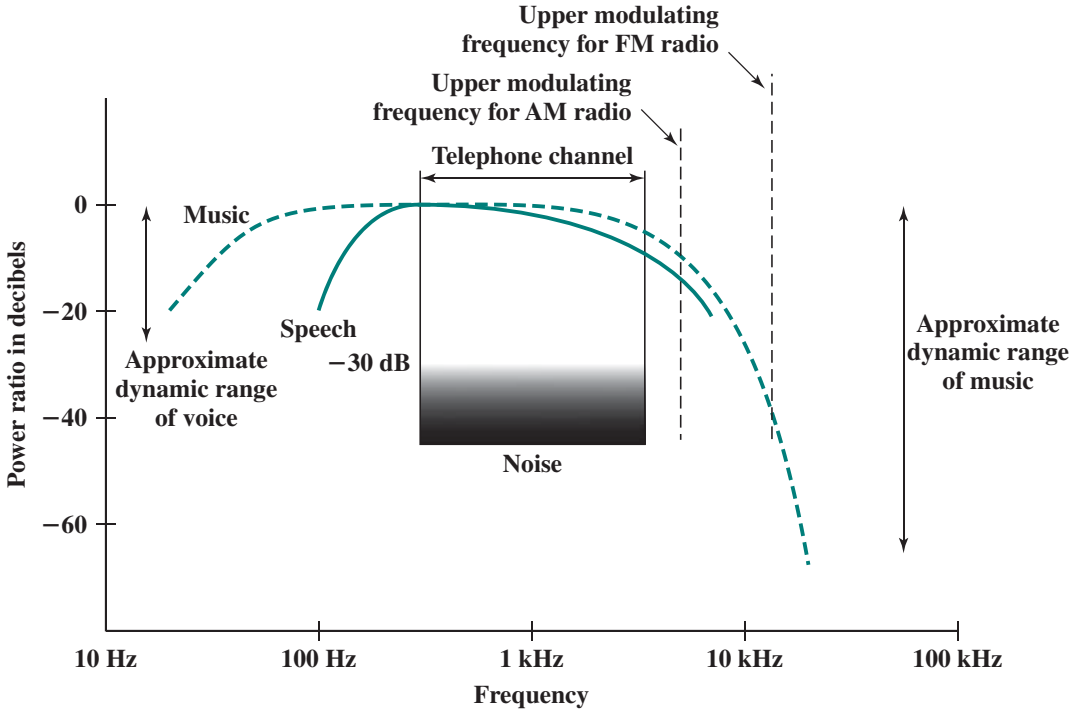


Figure 3.9 Acoustic Spectrum of Speech and Music [CARN99]

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 International Reference Alphabet (IRA).[8] 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.

Video transmission carries sequences of pictures in time. In essence, video makes use of a sequence of raster-scan images. Here it is easier to characterize the data in terms of the viewer's (destination's) television or computer display monitor rather than the original scene (source) that is recorded by the video camera.

Video can be captured by either analog or digital video recorders. The video that is captured can be transmitted using continuous (analog) or discrete (digital) signals, can be received by either analog or digital display devices, and can be stored in either analog or digital file formats.

The first televisions and computer monitors used cathode-ray-tube (CRT) technology. CRT monitors are inherently analog devices that use an electron gun to

[8]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 F provides a description and table of the IRA code.

"paint" pictures on the screen. The gun emits an electron beam that 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 the beam scans, the analog value changes. Thus the video image can be thought of as a time-varying analog signal.

The term *digital video* refers to the capture, manipulation, and storage of video in digital formats. Digital video cameras capture moving images digitally. In essence, this is done by taking a series of digital photographs, at a rate of at least 30 frames per second.

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 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 3.10 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 the three examples in the preceding subsection. For each example, we will describe the signal and estimate its bandwidth.

First, we consider **audio**, or acoustic, information. One form of acoustic information, of course, is human speech. This form of information is easily converted to an electromagnetic signal for transmission (Figure 3.11). 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.

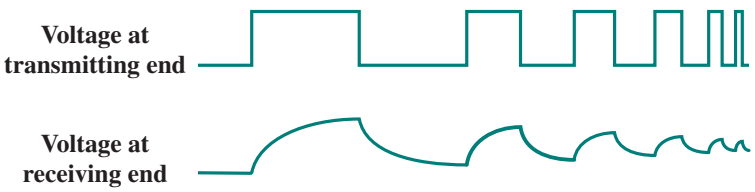
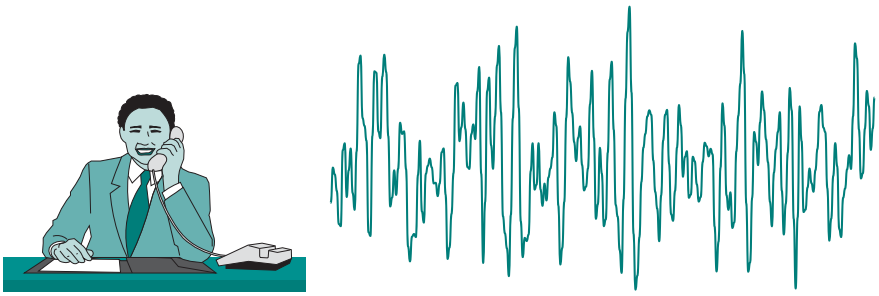


Figure 3.10 Attenuation of Digital Signals



In this graph of a typical analog voice 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 3.11 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. 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 $63.5\,\mu$s/line. Of the $63.5\,\mu$s, about $11\,\mu$s are allowed for horizontal retrace, leaving a total of $52.5\,\mu$s per video line.

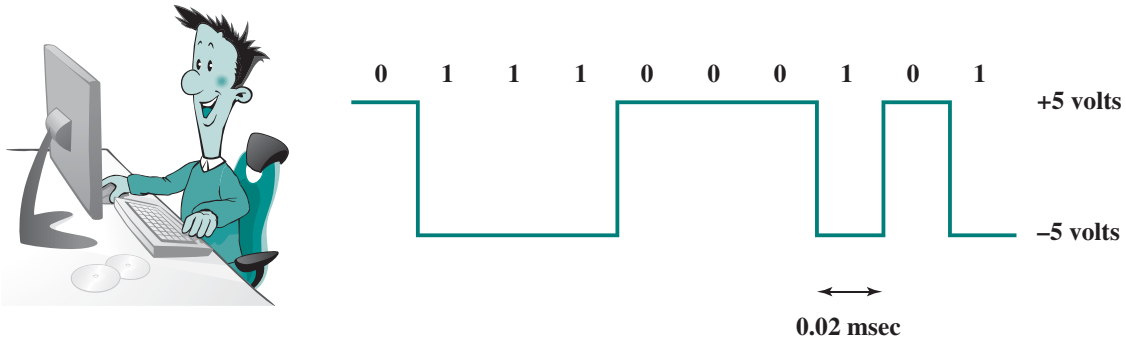
Now we are in a position to estimate the bandwidth required for the video signal. To do this we must estimate the upper (maximum) and lower (minimum) frequency of the band. We use the following reasoning to arrive at the maximum frequency: The maximum frequency would occur during the horizontal scan 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 μs, 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 MHz $-$ 0 $=$ 4 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 and then converted into digital voltage pulses for transmission, as illustrated in Figure 3.12. 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 nonreturn to zero (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 with 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



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 msec, giving a data rate of 50,000 bits per second (50 kbps).

Figure 3.12 Conversion of PC Input to Digital Signal

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 3.13 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.

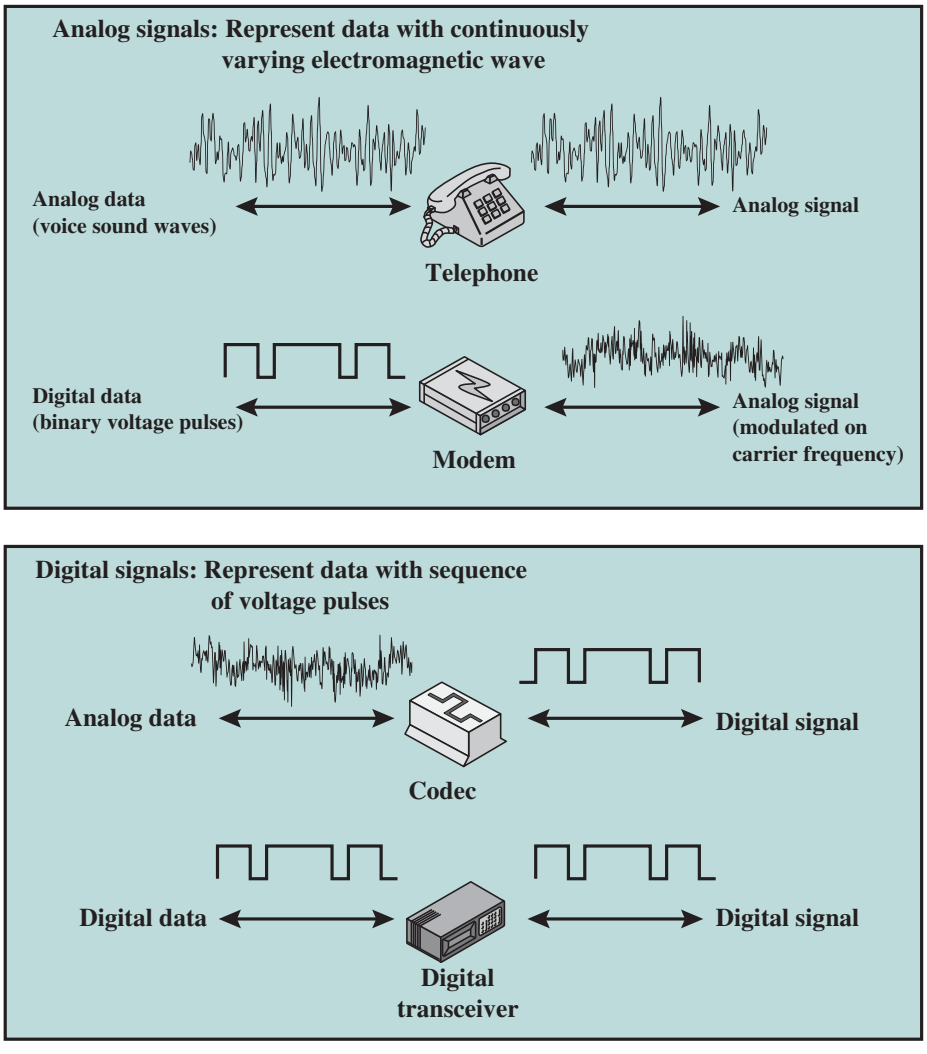


Figure 3.13 Analog and Digital Signaling of Analog and Digital Data

Thus, Figure 3.13 suggests that data may be encoded into signals in a variety of ways. We return to this topic in Chapter 5.

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. 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, by contrast, assumes a binary content to the signal. A digital signal can be transmitted only to a limited distance before attenuation, noise, and other impairments endanger the integrity of the data. To achieve greater

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 desired 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.

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 the following:

- **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.

Asynchronous and Synchronous Transmission

The reception of digital data involves sampling the incoming signal once per bit time to determine the binary value. One of the difficulties encountered in such a process is that various transmission impairments will corrupt the signal so that occasional errors will occur. This problem is compounded by a timing difficulty: In order for the receiver to sample the incoming bits properly, it must know the arrival time and duration of each bit that it receives.

Suppose that the sender simply transmits a stream of data bits. The sender has a clock that governs the timing of the transmitted bits. For example, if data are to be transmitted at 1 million bits per second (1 Mbps), then one bit will be transmitted every $1/10^6 = 1$ microsecond (μs), as measured by the sender's clock. Typically, the receiver will attempt to sample the medium at the center of each bit time. The receiver will time its samples at intervals of one bit time. In our example, the sampling would occur once every 1 μs. If the receiver times its samples based on its own clock, then there will be a problem if the transmitter's and receiver's clocks are not

precisely aligned. If there is a drift of 1% (the receiver's clock is 1% faster or slower than the transmitter's clock), then the first sampling will be 0.01 of a bit time (0.01 μs) away from the center of the bit (center of bit is 0.5 μs from beginning and end of bit). After 50 or more samples, the receiver may be in error because it is sampling in the wrong bit time ($50'.01 = 0.5\,\mu$s). For smaller timing differences, the error would occur later, but eventually the receiver will be out of step with the transmitter if the transmitter sends a sufficiently long stream of bits and if no steps are taken to synchronize the transmitter and receiver.

Two approaches are common for achieving the desired synchronization. The first is called, oddly enough, **asynchronous transmission**. The strategy with this scheme is to avoid the timing problem by not sending long, uninterrupted streams of bits. Instead, data are transmitted one character at a time, where each character is 5 to 8 bits in length. Timing or synchronization must only be maintained within each character; the receiver has the opportunity to resynchronize at the beginning of each new character.

With **synchronous transmission**, a block of bits is transmitted in a steady stream without start and stop codes. The block may be many bits in length. To prevent timing drift between transmitter and receiver, their clocks must somehow be synchronized. One possibility is to provide a separate clock line between transmitter and receiver. One side (transmitter or receiver) pulses the line regularly with one short pulse per bit time. The other side uses these regular pulses as a clock. This technique works well over short distances, but over longer distances the clock pulses are subject to the same impairments as the data signal, and timing errors can occur. The other alternative is to embed the clocking information in the data signal. For digital signals, this can be accomplished with Manchester or differential Manchester encoding. For analog signals, a number of techniques can be used; for example, the carrier frequency itself can be used to synchronize the receiver based on the phase of the carrier.

With synchronous transmission, there is another level of synchronization required to allow the receiver to determine the beginning and end of a block of data. To achieve this, each block begins with a *preamble* bit pattern and generally ends with a *postamble* bit pattern. In addition, other bits are added to the block that conveys control information used in the data link control procedures. The data plus preamble, postamble, and control information are called a **frame**. The exact format of the frame depends on which data link control procedure is being used.

Appendix D contains more detail on this topic.

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 introduce various random modifications that 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 (e.g., twisted-pair wire, optical fiber), this reduction in strength, or attenuation, is generally exponential and thus is typically expressed as a constant number of decibels per unit distance.[9] For unguided media (wireless transmission), attenuation is a more complex function of distance and the makeup of the atmosphere. Attenuation introduces three considerations for the transmission engineer.

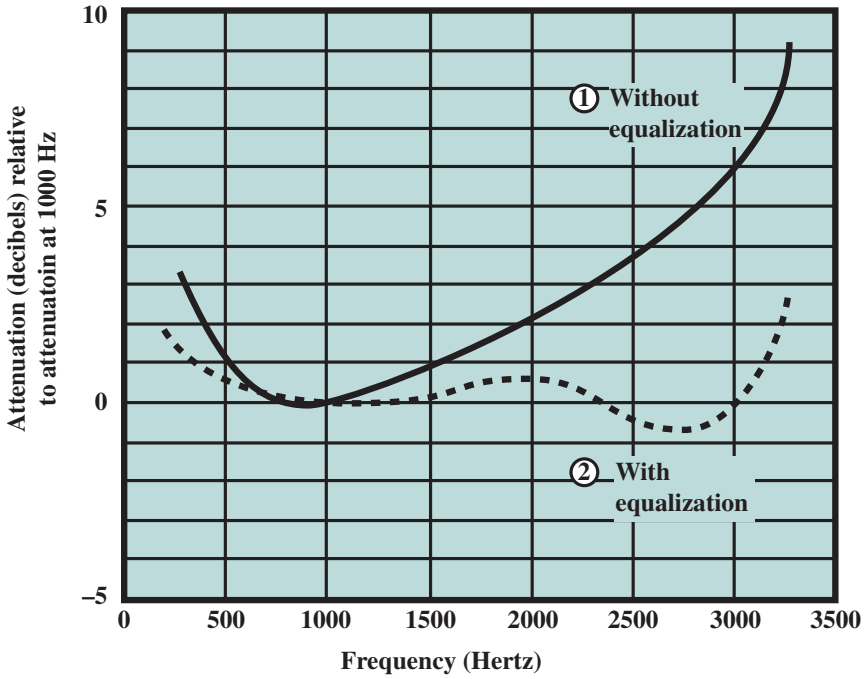
1. A received signal must have sufficient strength so that the electronic circuitry in the receiver can detect and interpret the signal.
2. The signal must maintain a level sufficiently higher than noise to be received without error.
3. Attenuation is greater at higher frequencies, and this causes distortion.

The first and second considerations 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 so great that repeaters or amplifiers must be installed at regular intervals to boost the signal. These problems are more complex for multipoint lines where the distance from transmitter to receiver is variable.

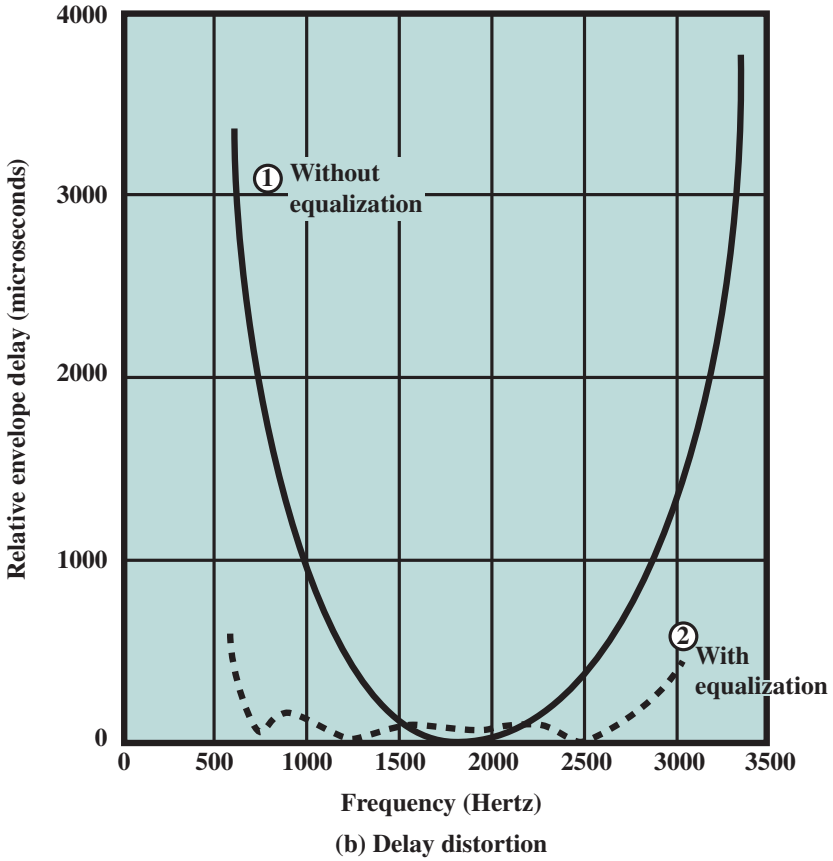
The third consideration, known as attenuation distortion, is particularly noticeable for analog signals. Because attenuation is different for different frequencies, and the signal is made up of a number of components at different frequencies, the received signal is not only reduced in strength but is also distorted. 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 3.14a, 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

[9]Standards documents generally use the term *insertion loss* when referring to losses associated with cabling, because it is more descriptive of a loss caused by the insertion of a link between transmitter and receiver. Because it remains a more familiar term, we generally use *attenuation* in this book.



(a) Attenuation



(b) Delay distortion

Figure 3.14 Attenuation and Delay Distortion Curves for a Voice Channel

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[10]

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

The solid line in Figure 3.14a 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 is a phenomenon that occurs in transmission cables (such as twisted pair, coaxial cable, and optical fiber); it does not occur when signals are transmitted through the air by means of antennas. Delay distortion is caused by the fact that the velocity of propagation of a signal through a cable is different for different frequencies. For a signal with a given bandwidth, the velocity tends to be highest near the center frequency of the signal and to fall off toward the two edges of the band. Thus, various components of a signal will arrive at the receiver at different times.

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 1 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.

Equalizing techniques can also be used for delay distortion. Again using a leased telephone line as an example, Figure 3.14b 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.

---

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

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[11]

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

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

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

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,

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

[11] 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 $T - 273.15$.

EXAMPLE 3.4 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, if two signals, one at 4000 Hz and one at 8000 Hz, share the same transmission facility, they might produce energy at 12,000 Hz. This noise could interfere with an intended signal at 12,000 Hz.

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 communication. For example, a sharp spike of energy of 0.01s duration would not destroy any voice data but would wash out about 560 bits of digital data being transmitted at 56 kbps. Figure 3.15 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.

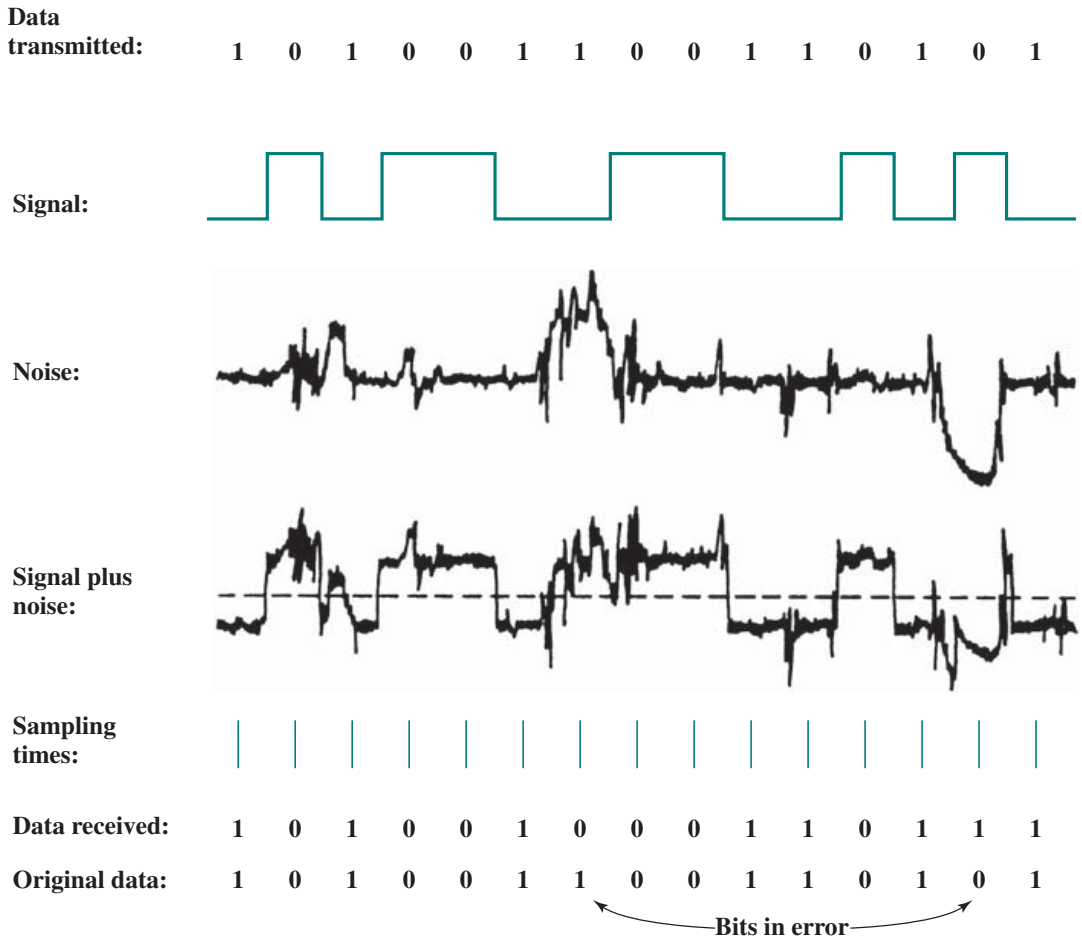


Figure 3.15 Effect of Noise on a Digital Signal

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 G.

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 1 bit. For example, if four possible voltage levels are used as signals, then each signal element can represent 2 bits. With multilevel signaling, the Nyquist formulation becomes

$$C = 2\mathrm{B}\log_2 M$$

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 3.5 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 1 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 3.15 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 [SHAN48]. 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),[12] 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:

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

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

$$C = B \log_2 (1 + \text{SNR}) \quad \textbf{(3.1)}$$

where C is the capacity of the channel in bits per second and B is the bandwidth of the channel in hertz.[13] 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 and delay distortion. Even in 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.

---

[12]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 SNR_{dB} or $(\text{S/N})_{\text{dB}}$. Others use just SNR or S/N to mean the dB quantity. This text uses SNR and SNR_{dB}.

[13]An intuitive proof of Shannon's equation is contained in a document at box.com/dcc10e.

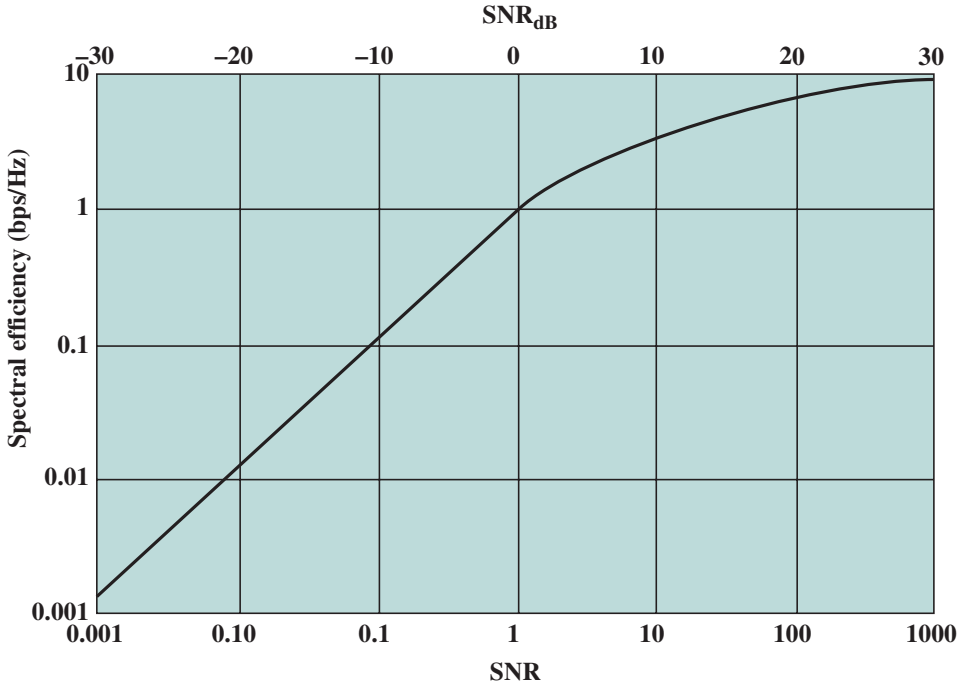


Figure 3.16 Spectral Efficiency versus SNR

We define the **spectral efficiency**, also called **bandwidth efficiency**, of a digital transmission as the number of bits per second of data that can be supported by each hertz of bandwidth. The theoretical maximum spectral efficiency can be expressed using Equation (3.1) by moving the bandwidth B to the left-hand side, resulting in $C/B = \log_2(1 + \text{SNR})$. C/B has the dimensions bps/Hz. Figure 3.16 shows the results on a log/log scale. At SNR = 1, we have $C/B = 1$. For SNR < 1 (signal power is less than noise power), the plot is linear; above SNR = 1, the plot flattens but continues to increase with increasing SNR.

We can make several observations about Figure 3.16. Below 0 dB SNR, noise is the dominant factor in the capacity of a channel. Shannon's theorem shows that communications is possible in this region, but at a relatively low data rate, a rate that is reduced in proportion to the SNR (on a log/log scale). In the region of at least 6 dB above 0 dB SNR, noise is no longer the limiting factor in communications speed. In this region, there is little ambiguity in a signal's relative amplitude and phase, and achieving a high-channel capacity depends on the design of the signal, including such factors as modulation type and coding.

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, the effects of nonlinearities in the system also increase, 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 3.6 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 $\text{SNR}_{\text{dB}} = 24$ dB. Then

$$\begin{aligned} B &= 4\ \text{MHz} - 3\ \text{MHz} = 1\ \text{MHz} \\ \text{SNR}_{\text{dB}} &= 24\,\text{dB} = 10\log_{10}(\text{SNR}) \\ \text{SNR} &= 251 \end{aligned}$$

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

$$\begin{aligned} C &= 2B\log_2 M \\ 8 \times 10^6 &= 2 \times (10^6) \times \log_2 M \\ 4 &= \log_2 M \\ M &= 16 \end{aligned}$$

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 1 bit. The data rate R is just $R = 1/T_b$. Thus

$$\frac{E_b}{N_0} = \frac{S/R}{N_0} = \frac{S}{\text{k}TR}$$

or, in decibel notation,

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

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 3.15. 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 2 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 3.7 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} (1 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

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

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

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

Another equation of interest relates $E\text{b}/N_0$ 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

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

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

EXAMPLE 3.8 Suppose we want to find the minimum $E\text{b}/N_0$ required to achieve a spectral efficiency of 6 bps/Hz. Then $E_b/N_0 = (1/6)(2^6 - 1) = 10.5 = 10.21$ dB.

3.5 RECOMMENDED READING

There are many books that cover the fundamentals of analog and digital transmission. [COUC13] is quite thorough. Other good reference works are [FREE05], which includes some of the examples used in this chapter, and [HAYK09].

COUC13 Couch, L. *Digital and Analog Communication Systems.* Upper Saddle River, NJ: Pearson, 2013.

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

HAYK09 Haykin, S. *Communication Systems.* New York: Wiley, 2009.

3.6 KEY TERMS, REVIEW QUESTIONS, AND PROBLEMS

Key Terms

absolute bandwidth	digital transmission	period
analog data	direct link	periodic signal
analog signal	discrete	point-to-point link
analog transmission	effective bandwidth	phase
aperiodic	error-free capacity	signal
asynchronous transmission	frame	signal-to-noise ratio (SNR)
attenuation	frequency	signaling
attenuation distortion	frequency domain	simplex
audio	full duplex	sinusoid
bandwidth	fundamental frequency	spectral efficiency
bandwidth efficiency	gain	spectrum
binary data	guided media	synchronous transmission
center frequency	half duplex	text
channel capacity	harmonic frequency	thermal noise
continuous	impulse noise	time domain
crosstalk	intermodulation noise	transmission
data	intersymbol interference	unguided media
dc component	loss	video
decibel (dB)	multipoint link	wavelength
delay distortion	noise	white noise
digital data	Nyquist bandwidth	wireless
digital signal	peak amplitude	

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 their simplest form:
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 $\cos a \cos b$.

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?

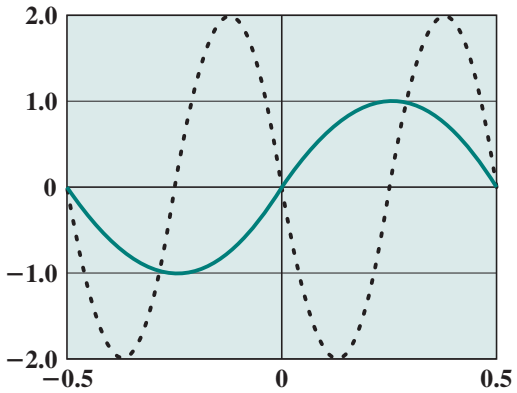


Figure 3.17 Figure for Problem 3.5

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. Commonly, medical digital radiology ultrasound studies consist of about 25 images extracted from a full-motion ultrasound examination. Each image consists of 512 by 512 pixels, each with 8 b of intensity information.

a. How many bits are there in the 25 images?

b. Ideally, however, doctors would like to use 512 × 512 8-bit frames at 30 fps (frames per second). Ignoring possible compression and overhead factors, what is the minimum channel capacity required to sustain this full-motion ultrasound?

c. Suppose each full-motion study consists of 25 s of frames. How many bytes of storage would be needed to store a single study in uncompressed form?

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 56 dB (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?

3.20. 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 as the maximum theoretical limit. How many signal levels are needed to achieve this data rate?

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 $N_0 = 0.1\,\mu$W/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 $E\text{b}/N_0$ for a link transmitting 2400 bps?

3.24. In a 1939 letter to Vannevar Bush, Claude Shannon said he was working on a theorem, which states that for any transmitter and receiver the length of an arbitrary message multiplied by its essential spectrum and divided by the distortion of the system is less than a certain constant times the time of transmission of the message multiplied by its essential spectrum width or, roughly speaking, it is impossible to reduce bandwidth times transmission time for a given distortion. Relate the theorem to Equation (3.1).

3.25. 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

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

3.27. 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.

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

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

where

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

Table 3.2 Decibel Values

Power Ratio	dB
1	0
2	3.01
10^1	10
10^2	20
10^3	30
10^4	40
10^5	50

Power Ratio	dB
0.5	−3.01
10^{-1}	−10
10^{-2}	−20
10^{-3}	−30
10^{-4}	−40
10^{-5}	−50
10^{-6}	−60

Table 3.2 shows the relationship between decibel values and powers of 10. The table also includes the decibel values for 2 and 1/2.

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 approximately 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 approximately 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:

$$L_{dB} = -10\log_{10}\frac{P_{out}}{P_{in}} = 10\log_{10}\frac{P_{in}}{P_{out}} \tag{3.3}$$

EXAMPLE 3.9 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

$$L_{dB} = 10\log(10/5) = 10(0.301) = 3.01\,\text{dB}.$$

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 approximately 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

P = power dissipated across resistance R
V = voltage across resistance R

Thus,

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

EXAMPLE 3.10 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}:

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

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:

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

EXAMPLE 3.11 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:

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

Note the following relationships:

$$+30\,\text{dBm} = 0\text{ dBW}$$
$$0\,\text{dBm} = -30\,\text{dBW}$$

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,

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

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