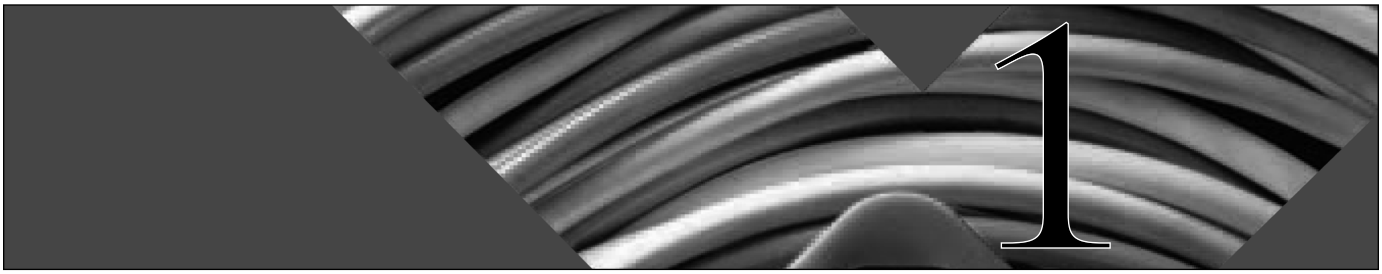
Chapter



INTRODUCTION

CHAPTER OBJECTIVES

- How communication systems work
- Frequency allocation and propagation characteristics
- Computer solutions (MATLAB)
- Information measure
- Coding performance

The subject of communication systems is immense. It is not possible to include all topics and keep one book of reasonable length. In this book, the topics are carefully selected to accentuate basic communication principles. For example, discussion is emphasized on the basic definition of instantaneous power (Chapter 2), average power (Chapter 2), and on the power of bandpass signals such as an AM radio signal (Chapter 4). Other basic concepts that are focused on are spectrum, signal-to-noise ratio for analog systems, and probability of bit error for digital systems. Moreover, the reader is motivated to appreciate these principles by the use of many practical applications. Often, practical applications are covered before the principles are fully developed. This provides "instant gratification" and motivates the reader to learn the basic principles well. The goal is to experience the joy of understanding how communication systems work and to develop an ability to design new communication systems.

This book is ideal for either a one-semester or a two-semester course. This book emphasizes basic material and applications that can be covered in a one-semester course, as well as the essential material that should be covered for a two-semester course. This emphasis means that the page count needs to be limited to around 750 pages. For a book with a larger page count, it is impossible to cover all that additional material, even in a two-semester course.

(Many schools are moving toward a basic one-course offering in communications.) Topics such as coding, wireless signal propagation, Wi MAX, and long-term evolution (LTE) of cellular systems are briefly covered in this book. In-depth coverage of important topics such as these should be done by additional courses with their own textbooks.

One major change for this eighth edition is the addition of more than 100 examples with solutions that are distributed throughout the chapters of the book. Students are always asking for more examples. Almost all of these new examples have a problem description that consists of only a few lines of text. The work for obtaining the solutions for these examples is done via MATLAB. These MATLAB solution files include the procedure for the solution (as described by comment lines in the MATLAB program), and then the results are computed and plotted. This presentation procedure has several advantages. First, the description for each example takes only a few lines in this textbook, so the book is not extended in length. Second, the student will have the experience of learning how to work with MATLAB (as demonstrated with the example solution). Clearly plotted results, which are better than hand calculations, will be given. The student can also vary the parameters in the MATLAB example to discover how the results will be affected.

What is a communication system? Moreover, what is electrical and computer engineering (ECE)? ECE is concerned with solving problems of two types: (1) production or transmission of electrical energy and (2) transmission or processing of information. *Communication systems are designed to transmit information.*

It is important to realize that communication systems and electric energy systems have markedly different sets of constraints. In electric energy systems, the waveforms are usually *known*, and one is concerned with designing the system for *minimum energy loss.*

In communication systems, the waveform present at the receiver (user) is *unknown* until after it is received—otherwise, no information would be transmitted, and there would be no need for the communication system. More information is communicated to the receiver when the user is "more surprised" by the message that was transmitted. That is, the transmission of information implies the communication of messages that are not known ahead of time (a priori).

Noise limits our ability to communicate. If there were no noise, we could communicate messages electronically to the outer limits of the universe by using an infinitely small amount of power. This has been intuitively obvious since the early days of radio. However, the theory that describes noise and the effect of noise on the transmission of information was not developed until the 1940s, by such persons as D. O. North [1943], S. O. Rice [1944], C. E. Shannon [1948], and N. Wiener [1949].

Communication systems are designed to transmit information bearing waveforms to the receiver. There are many possibilities for selecting waveforms to represent the information. For example, how does one select a waveform to represent the letter A in a typed message? Waveform selection depends on many factors. Some of these are bandwidth (frequency span) and center frequency of the waveform, waveform power or energy, the effect of noise on corrupting the information carried by the waveform, and the cost of generating the waveform at the transmitter and detecting the information at the receiver.

The book is divided into eight chapters and three appendices. Chapter 1 introduces some key concepts, such as the definition of information, and provides a method for evaluating the information capacity of a communication system. Chapter 2 covers the basic techniques for obtaining the spectrum bandwidth and power of waveforms. Baseband waveforms (which have frequencies near $f = 0$) are studied in Chapter 3, and bandpass waveforms (frequencies in some band not near $f = 0$) are examined in Chapters 4 and 5. The effect of noise on waveform selection is covered in Chapters 6 and 7. Case studies of wire and wireless communications, including personal communication systems (PCS) are emphasized in Chapter 8. The appendices include mathematical tables, a short course on probability and random variables, and an introduction to MATLAB. Standards for communications systems are included, as appropriate, in each chapter. The personal computer is used as a tool to plot waveforms, compute spectra of waveforms, and analyze and design communications systems.

In summary, communication systems are designed to transmit information. Communication system designers have four main concerns:

1. Selection of the information-bearing waveform
2. Bandwidth and power of the waveform
3. Effect of system noise on the received information
4. Cost of the system.

1–1 HISTORICAL PERSPECTIVE

A time chart showing the historical development of communications is given in Table 1–1. The reader is encouraged to spend some time studying this table to obtain an appreciation for the chronology of communications. Note that although the telephone was developed late in the 19th century, the first transatlantic telephone cable was not completed until 1954. Previous to that date, transatlantic calls were handled via shortwave radio. Similarly, although the British began television broadcasting in 1936, transatlantic television relay was not possible until 1962, when the *Telstar I* satellite was placed into orbit. Digital transmission systems—embodied by telegraph systems—were developed in the 1850s before analog systems—the telephone—in the 20th century. Now, digital transmission is again becoming the preferred technique.

TABLE 1–1 IMPORTANT DATES IN COMMUNICATIONS

Year	Event
Before 3000 B.C.	Egyptians develop a picture language called *hieroglyphics.*
A.D. 800	Arabs adopt our present number system from India.
1440	Johannes Gutenberg invents movable metal type.
1752	Benjamin Franklin's kite shows that lightning is electricity.
1827	Georg Simon Ohm formulates his law ($I = E/R$).
1834	Carl F. Gauss and Ernst H. Weber build the electromagnetic telegraph.
1838	William F. Cooke and Sir Charles Wheatstone build the telegraph.
1844	Samuel F. B. Morse demonstrates the Baltimore, MD, and Washington, DC, telegraph line.
1850	Gustav Robert Kirchhoff first publishes his circuit laws.
1858	The first transatlantic cable is laid and fails after 26 days.
1864	James C. Maxwell predicts electromagnetic radiation.
1871	The Society of Telegraph Engineers is organized in London.
1876	Alexander Graham Bell develops and patents the telephone.

TABLE 1–1 (*cont.*)

Year	Event
1883	Thomas A. Edison discovers a flow of electrons in a vacuum, called the “Edison effect,” the foundation of the electron tube.
1884	The American Institute of Electrical Engineers (AIEE) is formed.
1887	Heinrich Hertz verifies Maxwell’s theory.
1889	The Institute of Electrical Engineers (IEE) forms from the Society of Telegraph Engineers in London.
1894	Oliver Lodge demonstrates wireless communication over a distance of 150 yards.
1900	Guglielmo Marconi transmits the first transatlantic wireless signal.
1905	Reginald Fessenden transmits speech and music by radio.
1906	Lee deForest invents the vacuum-tube triode amplifier.
1907	The Society of Wireless Telegraph Engineers is formed in the United States.
1909	The Wireless Institute is established in the United States.
1912	The Institute of Radio Engineers (IRE) is formed in the United States from the Society of Wireless Telegraph Engineers and the Wireless Institute.
1915	Bell System completes a U.S. transcontinental telephone line.
1918	Edwin H. Armstrong invents the superheterodyne receiver circuit.
1920	KDKA, Pittsburgh, PA, begins the first scheduled radio broadcasts.
1920	J. R. Carson applies sampling to communications.
1923	Vladimir K. Zworkykin devises the “iconoscope” television pickup tube.
1926	J. L. Baird (England) and C. F. Jenkins (United States) demonstrate television.
1927	The Federal Radio Commission is created in the United States.
1927	Harold Black develops the negative-feedback amplifier at Bell Laboratories.
1928	Philo T. Farnsworth demonstrates the first all-electronic television system.
1931	Teletypewriter service is initiated.
1933	Edwin H. Armstrong invents FM.
1934	The Federal Communication Commission (FCC) is created from the Federal Radio Commission in the United States.
1935	Robert A. Watson-Watt develops the first practical radar.
1936	The British Broadcasting Corporation (BBC) begins the first television broadcasts.
1937	Alex Reeves conceives pulse code modulation (PCM).
1941	John V. Atanasoff invents the digital computer at Iowa State College.
1941	The FCC authorizes television broadcasting in the United States.
1945	The ENIAC electronic digital computer is developed at the University of Pennsylvania by John W. Mauchly.
1947	Walter H. Brattain, John Bardeen, and William Shockley devise the transistor at Bell Laboratories.
1947	Steve O. Rice develops a statistical representation for noise at Bell Laboratories.
1948	Claude E. Shannon publishes his work on information theory.
1950	Time-division multiplexing is applied to telephony.
1950s	Microwave telephone and communication links are developed.
1953	NTSC color television is introduced in the United States.
1953	The first transatlantic telephone cable (36 voice channels) is laid.
1957	The first Earth satellite, *Sputnik I*, is launched by USSR.

TABLE 1–1 (*cont.*)

Year	Event
1958	A. L. Schawlow and C. H. Townes publish the principles of the laser.
1958	Jack Kilby of Texas Instruments builds the first germanium integrated circuit (IC).
1958	Robert Noyce of Fairchild produces the first silicon IC.
1961	Stereo FM broadcasts begin in the United States.
1962	The first active satellite, *Telstar I*, relays television signals between the United States and Europe.
1963	Bell System introduces the touch-tone phone.
1963	The Institute of Electrical and Electronic Engineers (IEEE) is formed by merger of the IRE and AIEE.
1963–66	Error-correction codes and adaptive equalization for high-speed error-free digital communications are developed.
1964	The electronic telephone switching system (No. 1 ESS) is placed into service.
1965	The first commercial communications satellite, *Early Bird*, is placed into service.
1968	Cable television systems are developed.
1971	Intel Corporation develops the first single-chip microprocessor, the 4004.
1972	Motorola demonstrates the cellular telephone to the FCC.
1976	Personal computers are developed.
1979	64-kb random access memory ushers in the era of very large-scale integrated (VLSI) circuits.
1980	Bell System FT3 fiber-optic communication is developed.
1980	Compact disk is developed by Philips and Sony.
1981	IBM PC is introduced.
1982	AT&T agrees to divest its 22 Bell System telephone companies.
1984	Macintosh computer is introduced by Apple.
1985	FAX machines become popular.
1989	Global positioning system (GPS) using satellites is developed.
1995	The Internet and the World Wide Web become popular.
2000–present	Era of digital signal processing with microprocessors, digital oscilloscopes, digitally tuned receivers, megaflop personal computers, spread spectrum systems, digital satellite systems, digital television (DTV), and personal communications systems (PCS).

1–2 DIGITAL AND ANALOG SOURCES AND SYSTEMS

Definition. A *digital information source* produces a finite set of possible messages.

A telephone touchtone pad is a good example of a digital source. There is a finite number of characters (messages) that can be emitted by this source.

Definition. An *analog information source* produces messages that are defined on a continuum.

A microphone is a good example of an analog source. The output voltage describes the information in the sound, and it is distributed over a continuous range of values.

Definition. A *digital communication system* transfers information from a digital source to the intended receiver (also called the sink).

Definition. An *analog communication system* transfers information from an analog source to the sink.

Strictly speaking, a *digital waveform* is defined as a function of time that can have only a discrete set of amplitude values. If the digital waveform is a binary waveform, only two values are allowed. An *analog waveform* is a function of time that has a continuous range of values.

An electronic *digital* communication system usually has voltage and current waveforms that have digital values; however, it *may* have analog waveforms. For example, the information from a binary source may be transmitted to the receiver by using a sine wave of 1,000 Hz to represent a binary 1 and a sine wave of 500 Hz to represent a binary 0. Here the digital source information is transmitted to the receiver by the use of analog waveforms, but the system is still called a digital communication system. From this viewpoint, we see that a *digital* communication engineer needs to know how to analyze analog circuits as well as digital circuits.

Digital communication has a number of advantages:

- Relatively inexpensive digital circuits may be used.
- Privacy is preserved by using data encryption.
- Greater dynamic range (the difference between the largest and smallest values) is possible.
- Data from voice, video, and data sources may be merged and transmitted over a common digital transmission system.
- In long-distance systems, noise does not accumulate from repeater to repeater.
- Errors in detected data may be small, even when there is a large amount of noise on the received signal.
- Errors may often be corrected by the use of coding.

Digital communication also has disadvantages:

- Generally, more bandwidth is required than that for analog systems.
- Synchronization is required.

The advantages of digital communication systems usually outweigh their disadvantages. Consequently, digital systems are becoming dominant.

1–3 DETERMINISTIC AND RANDOM WAVEFORMS

In communication systems, we are concerned with two broad classes of waveforms: deterministic and random (or stochastic).

Definition. A *deterministic waveform* can be modeled as a completely specified function of time.

For example, if

$$w(t) = A \cos(\omega_0 t + \varphi_0) \tag{1–1}$$

describes a waveform, where A, ω_0, and φ_0 are known constants, this waveform is said to be deterministic because, for any value of t, the value $w(t)$ can be evaluated. If any of the constants are unknown, then the value of $w(t)$ cannot be calculated, and consequently, $w(t)$ is not deterministic.

Definition. A *random waveform* (or stochastic waveform) cannot be completely specified as a function of time and must be modeled probabilistically.†

Here we are faced immediately with a dilemma when analyzing communication systems. We know that the waveforms that represent the source cannot be deterministic. For example, in a digital communication system, we might send information corresponding to any one of the letters of the English alphabet. Each letter might be represented by a deterministic waveform, but when we examine the waveform that is emitted from the source, we find that it is a random waveform because we do not know exactly which characters will be transmitted. Consequently, we really need to design the communication system by using a random signal waveform. Noise would also be described by a random waveform. This requires the use of probability and statistical concepts (covered in Chapters 6 and 7) that make the design and analysis procedure more complicated. However, if we represent the signal waveform by a "typical" deterministic waveform, we can obtain most, but not all, of the results we are seeking. That is the approach taken in the first five chapters of this book.

1–4 ORGANIZATION OF THE BOOK

Chapters 1 to 5 use a deterministic approach in analyzing communication systems. This approach allows the reader to grasp some important concepts without the complications of statistical analysis. It also allows the reader who is not familiar with statistics to obtain a basic understanding of communication systems. However, the important topic of performance of communication systems in the presence of noise cannot be analyzed without the use of statistics. These topics are covered in Chapters 6 and 7 and Appendix B.†† Chapter 8 gives practical case studies of wire and wireless communication systems.

This textbook is designed to be reader friendly. To aid the student, there are more than 100 examples distributed throughout the chapters of this book, most of which have a MATLAB solution obtained via electronic M files that can be downloaded free-of-charge from the website indicated at the end of this paragraph. In addition, there are study-aid problems with solutions at the end of each chapter. The personal computer (PC) is used to solve problems as appropriate. In addition, a *Student Solutions Manual* contains detailed solutions for over 100 selected (out of 550) end-of-the-chapter homework problems. The selected problems are marked with a ★. For the selected problems that have computer solutions, MATLAB solution files are available to the student. To download the free *Student Solutions Manual* and the MATLAB files, go to http://lcouch.us or to http://couch.ece.ufl.edu.

† A more complete definition of a random waveform, also called a *random process*, is given in Chapter 6.

†† Appendix B covers the topic of probability and random variables and is a complete chapter in itself. This allows the reader who has not had a course on this topic to learn the material before Chapters 6 and 7 are studied.

If needed, an errata list for this book will also be posted on this website.

The book is also useful as a reference source for mathematics (Appendix A), statistics (Appendix B and Chapter 6), and MATLAB (Appendix C), and as a reference listing communication systems standards that have been adopted (Chapters 3, 4, 5, and 8).

Communications is an exciting area in which to work. The reader is urged to browse through Chapter 8, looking at case-study topics that are of special interest in both wireless and wire communication systems. To learn more about applied communication systems and examples of circuits that you can build, see or buy a recent edition of the *ARRL Handbook* [e.g., ARRL, 2010].

1–5 USE OF A PERSONAL COMPUTER AND MATLAB

This textbook is designed so that a PC may be used as a tool to plot waveforms; compute spectra (using the fast Fourier transform); evaluate integrals; and, in general, help the reader to understand, analyze, and design communication systems. MATLAB was chosen as the program language since it is very efficient at these tasks and a student version is available at a reasonable cost. For a brief summary of MATLAB programming concepts and instructions on running MATLAB, see Appendix C ("using MATLAB"). MATLAB files are provided for solving the Example problems, the Study-Aid problems, and selected end-of-the-chapter Homework problems. All of these files run on version 7.11 R2010b of MATLAB. These files are available for free downloading from the website indicated previously. See Appendix C for more details.

(Additional MATLAB files for all homework problems marked with the PC (💻) symbol are made available to the instructor and are included with the *Instructor Solutions Manual.*)

1–6 BLOCK DIAGRAM OF A COMMUNICATION SYSTEM

Communication systems may be described by the block diagram shown in Fig. 1–1. Regardless of the particular application, all communications systems involve three main subsystems: the *transmitter*, the *channel*, and the *receiver*. Throughout this book, we use the symbols as indicated in this diagram so that the reader will not be confused about where the signals are located in the overall system. The message from the source is represented by the information input waveform $m(t)$. The message delivered by the receiver is denoted by $\tilde{m}(t)$. The [~]

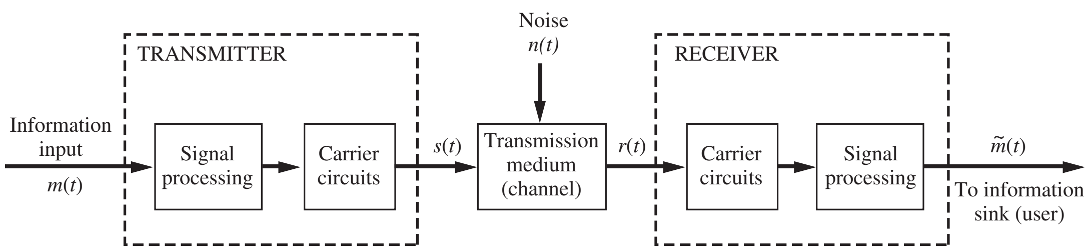


Figure 1–1 Communication system.

indicates that the message received may not be the same as that transmitted. That is, the message at the sink, $\tilde{m}(t)$, may be corrupted by noise in the channel, or there may be other impairments in the system, such as undesired filtering or undesired nonlinearities. The message information may be in analog or digital form, depending on the particular system, and it may represent audio, video, or some other type of information. In multiplexed systems, there may be multiple input and output message sources and sinks. The spectra (or frequencies) of $m(t)$ and $\tilde{m}(t)$ are concentrated about $f = 0$; consequently, they are said to be *baseband* signals.

The signal-processing block at the transmitter conditions the source for more efficient transmission. For example, in an analog system, the signal processor may be an analog low-pass filter that is used to restrict the bandwidth of $m(t)$. In a hybrid system, the signal processor may be an analog-to-digital converter (ADC), which produces a "digital word" that represents samples of the analog input signal (as described in Chapter 3 in the section on pulse code modulation). In this case, the ADC in the signal processor is providing *source coding* of the input signal. In addition, the signal processor may add parity bits to the digital word to provide *channel coding* so that error detection and correction can be used by the signal processor in the receiver to reduce or eliminate bit errors that are caused by noise in the channel. The signal at the output of the transmitter signal processor is a baseband signal, because it has frequencies concentrated near $f = 0$.

The transmitter carrier circuit converts the processed baseband signal into a frequency band that is appropriate for the transmission medium of the channel. For example, if the channel consists of a fiber-optic cable, the carrier circuits convert the baseband input (i.e., frequencies near $f = 0$) to light frequencies, and the transmitted signal, $s(t)$, is light. If the channel propagates baseband signals, no carrier circuits are needed, and $s(t)$ can be the output of the processing circuit at the transmitter. Carrier circuits are needed when the transmission channel is located in a band of frequencies around $f_c \gg 0$. (The subscript denotes "carrier" frequency.) In this case, $s(t)$ is said to be a *bandpass,* because it is designed to have frequencies located in a band about f_c. For example, an amplitude-modulated (AM) broadcasting station with an assigned frequency of 850 kHz has a carrier frequency of $f_c = 850$ kHz. The mapping of the baseband input information waveform $m(t)$ into the bandpass signal $s(t)$ is called *modulation.* [$m(t)$ is the audio signal in AM broadcasting.] In Chapter 4, it will be shown that any bandpass signal has the form

$$s(t) = R(t)\cos[\omega_c t + \theta(t)] \tag{1–2}$$

where $\omega_c = 2\pi f_c$. If $R(t) = 1$ and $\theta(t) = 0$, $s(t)$ would be a pure sinusoid of frequency $f = f_c$ with *zero* bandwidth. In the modulation process provided by the carrier circuits, the baseband input waveform $m(t)$ causes $R(t)$ or $\theta(t)$ or both to change as a function of $m(t)$. These fluctuations in $R(t)$ and $\theta(t)$ cause $s(t)$ to have a nonzero bandwidth that depends on the characteristics of $m(t)$ and on the mapping functions used to generate $R(t)$ and $\theta(t)$. In Chapter 5, practical examples of both digital and analog bandpass signaling are presented.

Channels may be classified into two categories: wire and wireless. Some examples of *wire* channels are twisted-pair telephone lines, coaxial cables, waveguides, and fiber-optic cables. Some typical *wireless* channels are air, vacuum, and seawater. Note that the general principles of digital and analog modulation apply to all types of channels, although channel characteristics may impose constraints that favor a particular type of signaling. In general, the channel medium attenuates the signal so that the noise of the channel or the noise introduced by an imperfect receiver causes the delivered information $\tilde{m}$ to be deteriorated from that of the source. The channel noise may arise from natural electrical disturbances (e.g., lightning) or from artificial sources, such as high-voltage transmission lines, ignition systems of cars, or switching circuits of a nearby digital computer. The channel may contain active amplifying devices, such as repeaters in telephone systems or satellite transponders in space communication systems.

The channel may also provide *multiple paths* between its input and output. This can be caused by the signal bouncing off of multiple reflectors. This *multipath* can be approximately described by two parameters—delay spread and Doppler spread. Delay spread is caused by multiple paths with varying lengths, which will cause a short transmitted pulse to be spread over time at the channel output because of the combination of received pulses with different delays from the multiple paths. Different motions of the various multipath reflectors cause the received pulses to have different Doppler frequency shifts so that there is a spread of Doppler frequency shifts on the components of the combined received signal. If the multipath reflectors move around slowly and, moreover, appear and disappear, the received signal will fade due to the individual received signals cancelling each other (when the composite received signal fades out). You have probably heard this fading effect on a distant AM radio station received at night. (The received night-time signals from distant AM stations are *skywave* signals as discussed in Sec. 1–8.)

The receiver takes the corrupted signal at the channel output and converts it to a baseband signal that can be handled by the receiver baseband processor. The baseband processor "cleans up" this signal and delivers an estimate of the source information $\tilde{m}(t)$ to the communication system output.

The goal is to design communication systems that transmit information to the receiver with as little deterioration as possible while satisfying design constraints, of allowable transmitted energy, allowable signal bandwidth, and cost. In digital systems, the measure of deterioration is usually taken to be the *probability of bit error* (P_e)—also called the *bit error rate* (BER)—of the delivered data $\tilde{m}$. In analog systems, the performance measure is usually taken to be the signal-to-noise ratio at the receiver output.

1–7 FREQUENCY ALLOCATIONS

Wireless communication systems often use the atmosphere for the transmission channel. Here, interference and propagation conditions are strongly dependent on the transmission frequency. Theoretically, any type of modulation (e.g., amplitude modulation, frequency modulation, single sideband, phase-shift keying, frequency-shift keying, etc.) could be used at any transmission frequency. However, to provide some semblance of order and to minimize interference, government regulations specify the modulation type, bandwidth, power, and type of information that a user can transmit over designated frequency bands.

Frequency assignments and technical standards are set internationally by the International Telecommunications Union (ITU). The ITU is a specialized agency of the United Nations, and the ITU administrative headquarters is located in Geneva, Switzerland, with a staff of about 700 persons (see http://www.itu.ch). This staff is responsible for administering the agreements that have been ratified by about 200 member nations of the ITU. The ITU is structured into three sectors. The Radiocommunication Sector (ITU-R) provides frequency assignments and is concerned with the efficient use of the radio frequency spectrum.

The Telecommunications Standardization Section (ITU-T) examines technical, operating, and tariff questions. It recommends worldwide standards for the public telecommunications network (PTN) and related radio systems. The Telecommunication Development Sector (ITU-D) provides technical assistance, especially for developing countries. This assistance encourages a full array of telecommunication services to be economically provided and integrated into the worldwide telecommunication system. Before 1992, the ITU was organized into two main sectors: the International Telegraph and Telephone Consultative Committee (CCITT) and the International Radio Consultative Committee (CCIR).

Each member nation of the ITU retains sovereignty over the spectral usage and standards adopted in its territory. However, each nation is expected to abide by the overall frequency plan and standards that are adopted by the ITU. Usually, each nation establishes an agency that is responsible for the administration of the radio frequency assignments within its borders. In the United States, the Federal Communications Commission (FCC) regulates and licenses radio systems for the general public and state and local government (see http://www.fcc.gov). In addition, the National Telecommunication and Information Administration (NTIA) is responsible for U.S. government and U.S. military frequency assignments. The international frequency assignments are divided into subbands by the FCC to accommodate 70 categories of services and 9 million transmitters. Table 1–2 gives a general listing of frequency bands, their common designations, typical propagation conditions, and typical services assigned to these bands.

TABLE 1–2 FREQUENCY BANDS

Frequency Band[a]	Designation	Propagation Characteristics	Typical Uses
3–30 kHz	Very low frequency (VLF)	Ground wave; low attenuation day and night; high atmospheric noise level	Long-range navigation; submarine communication
30–300 kHz	Low frequency (LF)	Similar to VLF, slightly less reliable; absorption in daytime	Long-range navigation and marine communication radio beacons
300–3000 kHz	Medium frequency (MF)	Ground wave and night sky wave; attenuation low at night and high in day; atmospheric noise	Maritime radio, direction finding, and AM broadcasting
3–30 MHz	High frequency (HF)	Ionospheric reflection varies with time of day, season, and frequency; low atmospheric noise at 30 MHz	Amateur radio; international broadcasting, military communication, long-distance aircraft and ship communication, telephone, telegraph, facsimile
30–300 MHz	Very high frequency (VHF)	Nearly line-of-sight (LOS) propagation, with scattering because of temperature inversions, cosmic noise	VHF television, FM two-way radio, AM aircraft communication, aircraft navigational aids

[a] kHz = 10^3 Hz; MHz = 10^6 Hz; GHz = 10^9 Hz.

TABLE 1–2 (*cont.*)

Frequency Band[a]	Designation	Propagation Characteristics	Typical Uses
0.3–3 GHz	Ultrahigh frequency (UHF)	LOS propagation, cosmic noise	UHF television, cellular telephone, navigational aids, radar, GPS, microwave links, personal communication systems
	Letter designation		
1.0–2.0	L		
2.0–4.0	S		
3–30 GHz	Superhigh frequency (SHF)	LOS propagation; rainfall attenuation above 10 GHz, atmospheric attenuation because of oxygen and water vapor, high water-vapor absorption at 22.2 GHz	Satellite communication, radar microwave links
	Letter designation		
2.0–4.0	S		
4.0–8.0	C		
8.0–12.0	X		
12.0–18.0	Ku		
18.0–27.0	K		
27.0–40.0	Ka		
26.5–40.0	R		
30–300 GHz	Extremely high frequency (EHF)	Same; high water-vapor absorption at 183 GHz and oxygen absorption at 60 and 119 GHz	Radar, satellite, experimental
	Letter designation		
27.0–40.0	Ka		
26.5–40.0	R		
33.0–50.0	Q		
40.0–75.0	V		
75.0–110.0	W		
110–300	mm (millimeter)		
10^3–10^7 GHz	Infrared, visible light, and ultraviolet	LOS propagation	Optical communications

[a] kHz $= 10^3$ Hz; MHz $= 10^6$ Hz; GHz $= 10^9$ Hz.

For a detailed chart of current frequency allocations in the United States see http://www.ntica. doc.gov/osmhome/allochrt.html.

1–8 PROPAGATION OF ELECTROMAGNETIC WAVES

The propagation characteristics of electromagnetic waves used in wireless channels are highly dependent on the frequency. This situation is shown in Table 1–2, where users are assigned frequencies that have the appropriate propagation characteristics for the coverage needed. The propagation characteristics are the result of changes in the radio-wave velocity as

M01_COUC5380_08_SE_C01.QXD 11/18/11 8:09 PM Page 13

a function of altitude and boundary conditions. The wave velocity is dependent on air temperature, air density, and levels of air ionization.

Ionization (i.e., free electrons) of the rarified air at high altitudes has a dominant effect on wave propagation in the medium-frequency (MF) and high-frequency (HF) bands. The ionization is caused by ultraviolet radiation from the sun, as well as cosmic rays. Consequently, the amount of ionization is a function of the time of day, season of the year, and activity of the sun (sunspots). This results in several layers of varying ionization density located at various heights surrounding the Earth.

The dominant ionized regions are D, E, F_1, and F_2 layers. The D layer is located closest to the Earth's surface at an altitude of about 45 or 55 miles. For $f > 300$ kHz, the D layer acts as a radio-frequency (RF) sponge to absorb (or attenuate) these radio waves. The attenuation is inversely proportional to frequency and becomes small for frequencies above 4 MHz. For $f < 300$ kHz, the D layer provides refraction (bending) of RF waves. The D layer is most pronounced during the daylight hours, with maximum ionization when the sun is overhead, and almost disappears at night. The E layer has a height of 65 to 75 miles, has maximum ionization around noon (local time), and practically disappears after sunset. It provides reflection of HF frequencies during the daylight hours. The F layer ranges in altitude between 90 and 250 miles. It ionizes rapidly at sunrise, reaches its peak ionization in early afternoon, and decays slowly after sunset. The F region splits into two layers, F_1 and F_2, during the day and combines into one layer at night. The F region is the most predominant medium in providing reflection of HF waves. As shown in Fig. 1–2, the electromagnetic spectrum may be divided into three broad bands that have one of three dominant propagation characteristics: ground wave, sky wave, and line of sight (LOS).

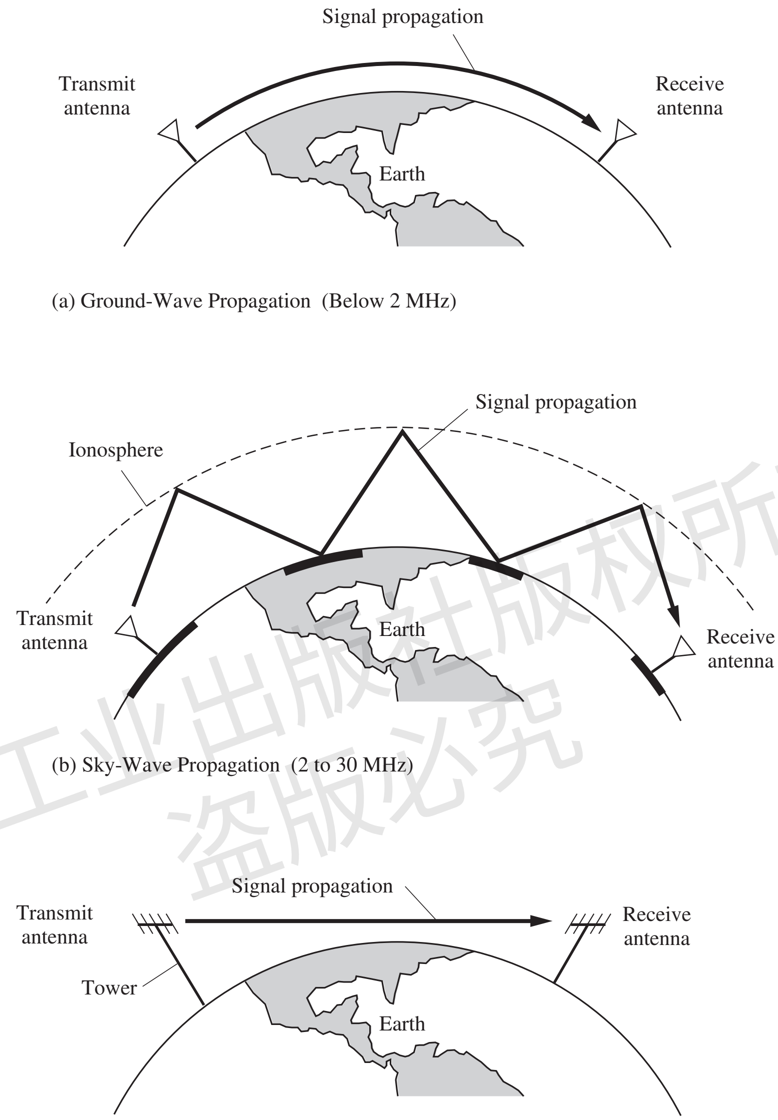
Ground-wave propagation is illustrated in Fig. 1–2a. It is the dominant mode of propagation for frequencies below 2 MHz. Here, the electromagnetic wave tends to follow the contour of the Earth. That is, diffraction of the wave causes it to propagate along the surface of the Earth. This is the propagation mode used in AM broadcasting, where the local coverage follows the Earth's contour and the signal propagates over the visual horizon. The following question is often asked: What is the lowest radio frequency that can be used? The answer is that the value of the lowest useful frequency depends on how long you want to make the antenna. For efficient radiation, the antenna needs to be longer than one-tenth of a wavelength. For example, for signaling with a carrier frequency of $f_c = 10$ kHz, the wavelength is

$$\lambda = \frac{c}{f_c}$$

$$\lambda = \frac{(3 \times 10^8 \text{ m/s})}{10^4} = 3 \times 10^4 \text{ m} \tag{1–3}$$

where c is the speed of light. (The formula $\lambda = c/f_c$ is distance $=$ velocity $\times$ time, where the time needed to traverse one wavelength is $t = 1/f_c$.) Thus, an antenna needs to be at least 3,000 m in length for efficient electromagnetic radiation at 10 kHz.

Sky-wave propagation is illustrated in Fig. 1–2b. It is the dominant mode of propagation in the 2- to 30-MHz frequency range. Here, long-distance coverage is obtained by reflecting the wave at the ionosphere, and at the Earth's boundaries. Actually, in the ionosphere the



(c) Line-of-Sight (LOS) Propagation (Above 30 MHz)

Figure 1–2 Propagation of radio frequencies.

waves are refracted (i.e., bent) gradually in an inverted U shape, because the index of refraction varies with altitude as the ionization density changes. The refraction index of the ionosphere is given by [Griffiths, 1987; Jordan and Balmain, 1968]

$$n = \sqrt{1 - \frac{81N}{f^2}} \tag{1–4}$$

where n is the refractive index, N is the free-electron density (number of electrons per cubic meter), and f is the frequency of the wave (in hertz). Typical N values range between 10^{10} and 10^{12}, depending on the time of day, the season, and the number of sunspots. In an ionized region $n < 1$ because $N > 0$, and outside the ionized region $n \approx 1$ because $N \approx 0$. In the ionized region, because $n < 1$, the waves will be bent according to Snell's law; viz,

$$n \sin \varphi_r = \sin \varphi_i \tag{1–5}$$

where φ_i is the angle of incidence (between the wave direction and vertical), measured just below the ionosphere, and φ_r is the angle of refraction for the wave (from vertical), measured in the ionosphere. Furthermore, the refraction index will vary with altitude within the ionosphere because N varies. For frequencies selected from the 2- to 30-MHz band, the refraction index will vary with altitude over the appropriate range so that the wave will be bent back to Earth. Consequently, the ionosphere acts as a reflector. The transmitting station will have coverage areas as indicated in Fig. 1–2b by heavy black lines along the Earth's surface. The coverage near the transmit antenna is due to the ground-wave mode, and the other coverage areas are due to sky wave. Notice that there are areas of no coverage along the Earth's surface between the transmit and receive antennas. The angle of reflection and the loss of signal at an ionospheric reflection point depend on the frequency, the time of day, the season of the year, and the sunspot activity [Jordan, 1985, Chap. 33].

During the daytime (at the ionospheric reflection points), the electron density will be high, so that $n < 1$. Consequently, sky waves from distant stations on the other side of the world will be heard on the shortwave bands. However, the D layer is also present during the day. This absorbs frequencies below 4 MHz.

This is the case for AM broadcast stations, where distant stations cannot be heard during the day, but at night the layer disappears, and distant AM stations can be heard via sky-wave propagation. In the United States, the FCC has designated some frequencies within the AM band as *clear channels* (as shown in Table 5–1). On these channels, only one or two high-power 50-kw stations are assigned to operate at night, along with a few low-power stations. Since these channels are relatively free of interfering stations, night sky-wave signals of the dominant 50-kw station can often be heard at distances up to 800 miles from the station. For example, some clear-channel 50-kw stations are WSM, Nashville, on 650 kHz; WCCO, Minneapolis, on 830 kHz; and WHO, Des Moines, on 1040 kHz. Actually, these "clear channels" are not so clear anymore since additional stations have been licensed for these channels as years have passed.

Sky-wave propagation is caused primarily by reflection from the F layer (90 to 250 miles in altitude). Because of this layer, international broadcast stations in the HF band can be heard from the other side of the world almost anytime during the day or night.

LOS propagation (illustrated in Fig. 1–2c) is the dominant mode for frequencies above 30 MHz. Here, the electromagnetic wave propagates in a straight line. In this case, $f^2 \gg 81N$, so that $n \approx 1$, and there is very little refraction by the ionosphere. In fact, the signal will propagate *through* the ionosphere. This property is used for satellite communications.

The LOS mode has the disadvantage that, for communication between two terrestrial (Earth) stations, the signal path has to be above the horizon. Otherwise, the Earth will block the LOS path. Thus, antennas need to be placed on tall towers so that the receiver antenna can

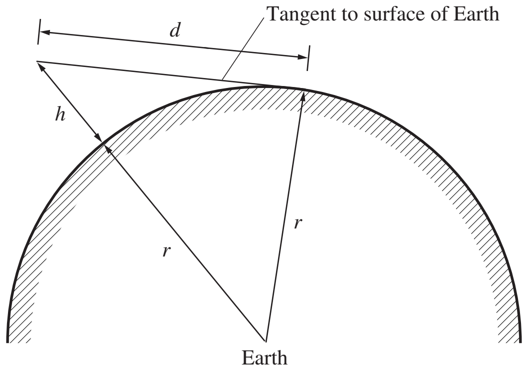


Figure 1–3 Calculation of distance to horizon.

"see" the transmitting antenna. A formula for the distance to the horizon, d, as a function of antenna height can be easily obtained by the use of Fig. 1–3. From this figure,

$$d^2 + r^2 = (r + h)^2$$

or

$$d^2 = 2rh + h^2$$

where r is the radius of the Earth and h is the height of the antenna above the Earth's surface. In this application, h^2 is negligible with respect to $2rh$. The radius of the Earth is 3,960 statute miles. However, at LOS radio frequencies the effective Earth radius† is $\frac{4}{3}$ (3,960) miles. Thus, the distance to the radio horizon is

$$d = \sqrt{2h} \text{ miles} \tag{1–6}$$

where conversion factors have been used so that h is the antenna height measured in feet and d is in statute miles. For example, television stations have assigned frequencies above 30 MHz in the VHF or UHF range (see Table 1–2), and the fringe-area coverage of high-power stations is limited by the LOS radio horizon. For a television station with a 1,000-ft tower, d is 44.7 miles. For a fringe-area viewer who has an antenna height of 30 ft, d is 7.75 miles. Thus, for these transmitting and receiving heights, the television station would have fringe-area coverage out to a radius of 44.7 + 7.75 = 52.5 miles around the transmitting tower.

Example 1–1 LINE OF SIGHT

Plot a graph describing the LOS distance in miles for transmitting antennas with heights from 0 ft up to 1000 ft. Assume that the receiving antenna is 5 ft above the ground. For the solution, see and run the MATLAB file Example1_1.m.

---

† The refractive index of the atmosphere decreases slightly with height, which causes some bending of radio rays. This effect may be included in LOS calculations by using an effective Earth radius that is four-thirds of the actual radius.

In addition to the LOS propagation mode, it is possible to have *ionospheric scatter propagation.* This mode occurs over the frequency range of 30 to 60 MHz, when the radio frequency signal is scattered because of irregularities in the refractive index of the lower ionosphere (about 50 miles above the Earth's surface). Because of the scattering, communications can be carried out over path lengths of 1,000 miles, even though that is beyond the LOS distance. Similarly, *tropospheric scattering* (within 10 miles above the Earth's surface) can propagate radio frequency signals that are in the 40-MHz to 4-GHz range over paths of several hundred miles.

For more technical details about radio-wave propagation, the reader is referred to textbooks that include chapters on ground-wave and sky-wave propagation [Griffiths, 1987; Jordan and Balmain, 1968] and on wireless cellular-signal propagation [Rappaport, 2002]. A very readable description of this topic is also found in the ARRL handbook [ARRL, 2010]. Personal computer programs that predict sky-wave propagation conditions, such as VOACAP [ARRL, 2010] are available.

1–9 INFORMATION MEASURE

As we have seen, the purpose of communication systems is to transmit information from a source to a receiver. However, what exactly is information, and how do we measure it? We know qualitatively that it is related to the surprise that is experienced when we receive the message. For example, the message "The ocean has been destroyed by a nuclear explosion" contains more information than the message "It is raining today."

Definition. The *information* sent from a digital source when the jth message is transmitted is given by

$$I_j = \log_2\left(\frac{1}{P_j}\right)\text{bits} \tag{1–7a}$$

where P_j is the probability of transmitting the jth message.[†]

From this definition, we see that messages that are less likely to occur (smaller value for P_j) provide more information (larger value of I_j). We also observe that the information measure depends on only the likelihood of sending the message and does not depend on possible interpretation of the content as to whether or not it makes sense.

The base of the logarithm determines the units used for the information measure. Thus, for units of "bits," the base 2 logarithm is used. If the natural logarithm is used, the units are "nats" and for base 10 logarithms, the unit is the "hartley," named after R. V. Hartley, who first suggested using the logarithm measure in 1928 [Hartley, 1948].

In this section, the term *bit* denotes a unit of information as defined by Eq. (1–7a). In later sections, particularly in Chapter 3, *bit* is also used to denote a unit of binary data.These two different meanings for the word *bit* should not be confused. Some authors use *binit* to denote units of data and use *bit* exclusively to denote units of information. However, most engineers use the same word (*bit*) to denote both kinds of units, with the particular meaning understood from the context in which the word is used. This book follows that industry custom.

[†] The definition of probability is given in Appendix B.

M01_COUC5380_08_SE_C01.QXD 11/18/11 8:09 PM Page 18

For ease of evaluating I_j on a calculator, Eq. (1–7a) can be written in terms of the base 10 logarithm or the natural logarithm:

$$I_j = -\frac{1}{\log_{10} 2} \log_{10} P_j = -\frac{1}{\ln 2} \ln P_j \tag{1–7b}$$

In general, the information content will vary from message to message because the P_j's will not be equal. Consequently, we need an average information measure for the source, considering all the possible messages we can send.

Definition. The *average information* measure of a digital source is

$$H = \sum_{j=1}^{m} P_j I_j = \sum_{j=1}^{m} P_j \log_2\left(\frac{1}{P_j}\right) \text{bits} \tag{1–8}$$

where m is the number of possible different source messages and P_j is the probability of sending the jth message (m is finite because a digital source is assumed). The average information is called *entropy.*

Example 1–2 ENTROPY

For a case of a binary problem, examine how the entropy changes as a function of P_1 where P_1 is the probability of obtaining a binary "1" and it ranges from 0 to 1. For the solution, see and run the MATLAB file Example1_2.m.

Example 1–3 EVALUATION OF INFORMATION AND ENTROPY

Find the information content of a message that consists of a digital word 12 digits long in which each digit may take on one of four possible levels. The probability of sending any of the four levels is assumed to be equal, and the level in any digit does not depend on the values taken on by previous digits.

In a string of 12 symbols (digits), where each symbol consists of one of four levels, there are $4 \cdot 4 \ldots 4 = 4^{12}$ different combinations (words) that can be obtained. Because each level is equally likely, all the different words are equally likely. Thus,

$$p_j = \frac{1}{4^{12}} = \left(\frac{1}{4}\right)^{12}$$

or

$$I_j = \log_2\left(\frac{1}{\left(\frac{1}{4}\right)^{12}}\right) = 12 \log_2(4) = 24 \text{ bits}$$

In this example, we see that the information content in every one of the possible messages equals 24 bits. Thus, the average information H is 24 bits.

Suppose that only two levels (binary) had been allowed for each digit and that all the words were equally likely. Then the information would be $I_j = 12$ bits for the binary words, and the average information would be $H = 12$ bits. Here, all the 12-bit words gave 12 bits of information, because the words were equally likely. If they had not been equally likely, some of the 12-bit words would contain more than 12 bits of information and some would contain less, and the average information would have been less than 12 bits. For example, if half of the 12-bit words (2,048 of the possible 4,096) have probability of occurence of $P_j = 10^{-5}$ for each of these words (with a corresponding $I_j = 16.61$ bits) and the other half have $P_j = 4.78 \times 10^{-4}$ (for a corresponding $I_j = 11.03$ bits), then the average information is $H = 11.14$ bits.

The rate of information is also important.

Definition. The *source rate* is given by

$$R = \frac{H}{T} \text{ bits/s} \tag{1–9}$$

where H is evaluated by using Eq. (1–8) and T is the time required to send a message.

The definitions previously given apply to digital sources. Results for analog sources can be approximated by digital sources with as much accuracy as we desire.

1–10 CHANNEL CAPACITY AND IDEAL COMMUNICATION SYSTEMS

Many criteria can be used to measure the effectiveness of a communication system to see if it is ideal or perfect. For digital systems, the optimum system might be defined as the system that minimizes the probability of bit error at the system output subject to constraints on transmitted energy and channel bandwidth. Thus, bit error and signal bandwidth are of prime importance and are covered in subsequent chapters. This raises the following question: Is it possible to invent a system with no bit error at the output even when we have noise introduced into the channel? This question was answered by Claude Shannon in 1948–1949 [Wyner and Shamai, 1998; Shannon, 1948, 1949]. The answer is yes, under certain assumptions. Shannon showed that (for the case of signal plus white Gaussian noise) a channel capacity C (bits/s) could be calculated such that if the rate of information R (bits/s) was less than C, the probability of error would approach zero. The equation for C is

$$C = B \log_2\left(1 + \frac{S}{N}\right) \tag{1–10}$$

where B is the channel bandwidth in hertz (Hz) and S/N is the signal-to-noise power ratio (watts/watts, not dB) at the input to the digital receiver. Shannon does not tell us how to build this system, but he proves that it is theoretically possible to have such a system. Thus, Shannon gives us a theoretical performance bound that we can strive to achieve with practical communication systems. Systems that approach this bound usually incorporate error-correction coding.

Example 1–4 CHANNEL CAPACITY

For a case of a telephone line with a bandwith of 3,300Hz, plot the channel capacity of the telephone line as a function of the S/N over a range of 0–60 dB. For the solution, see and run the MATLAB file Example1_4.m.

In analog systems, the optimum system might be defined as the one that achieves the largest signal-to-noise ratio at the receiver output, subject to design constraints such as channel bandwidth and transmitted power. Here, the evaluation of the output signal-to-noise ratio is of prime importance. We might ask the question, Is it possible to design a system with infinite signal-to-noise ratio at the output when noise is introduced by the channel? The answer is no. The performance of practical analog systems with respect to that of Shannon's ideal system is illustrated in Chapter 7. (See Fig. 7–27.)

Other fundamental limits for digital signaling were discovered by Nyquist in 1924 and Hartley in 1928. Nyquist showed that if a pulse represents one bit of data, noninterfering pulses could be sent over a channel no faster than $2B$ pulses/s, where B is the channel bandwidth in hertz. This is now known as the dimensionality theorem and is discussed in Chapter 2. Hartley generalized Nyquist's result for the case of multilevel pulse signaling, as discussed in Chapters 3 and 5.

The following section describes the improvement that can be obtained in digital systems when coding is used and how these coded systems compare with Shannon's ideal system.

1–11 CODING

If the data at the output of a digital communication system have errors that are too frequent for the desired use, the errors can often be reduced by the use of either of two main techniques:

- Automatic repeat request (ARQ)
- Forward error correction (FEC)

In an ARQ system, when a receiver circuit detects parity errors in a block of data, it requests that the data block be retransmitted. In an FEC system, the transmitted data are encoded so that the receiver can correct, as well as detect, errors. These procedures are also classified as *channel coding* because they are used to correct errors caused by channel noise. This is different from *source coding*, described in Chapter 3, where the purpose of the coding is to extract the essential information from the source and encode it into digital form so that it can be efficiently stored or transmitted using digital techniques.

The choice between using the ARQ or the FEC technique depends on the particular application. ARQ is often used in computer communication systems because it is relatively inexpensive to implement and there is usually a duplex (two-way) channel so that the receiving end can transmit back an acknowledgment (ACK) for correctly received data or a request for retransmission (NAC) when the data are received in error. FEC techniques are used to correct errors on simplex (one-way) channels, where returning of an ACK/NAC indicator (required for the ARQ technique) is not feasible. FEC is preferred on systems with large transmission delays, because if the ARQ technique were used, the effective data rate would be small; the transmitter would have long idle periods while waiting for the ACK/NAC indicator, which is retarded by the long transmission delay. We concentrate on FEC techniques in the remainder of this section.

Communication systems with FEC are illustrated in Fig. 1–4, where encoding and decoding blocks have been designated. Coding involves adding extra (redundant) bits to the data stream so that the decoder can reduce or correct errors at the output of the receiver.

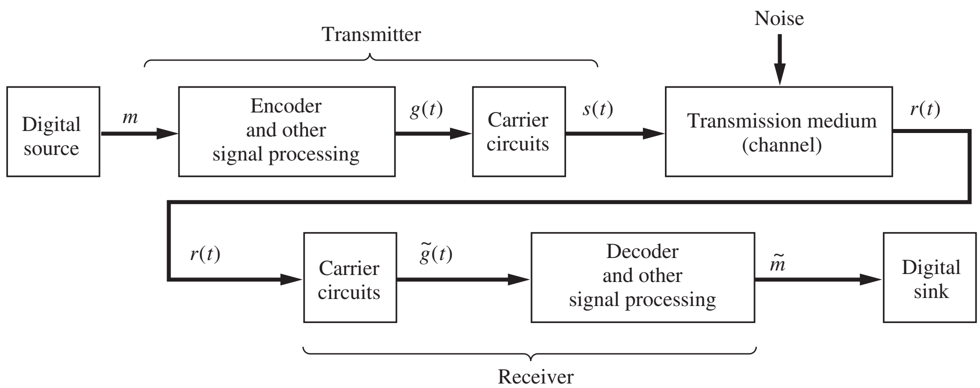


Figure 1–4 General digital communication system.

However, these extra bits have the disadvantage of increasing the data rate (bits/s) and, consequently, increasing the bandwidth of the encoded signal.

Codes may be classified into two broad categories:

- *Block codes.* A block code is a mapping of k input binary symbols into n output binary symbols. Consequently, the block coder is a *memoryless* device. Because $n > k$, the code can be selected to provide redundancy, such as *parity bits*, which are used by the decoder to provide some error detection and error correction. The codes are denoted by (n, k), where the code rate $R^{\dagger}$ is defined by $R = k/n$. Practical values of R range from $\frac{1}{4}$ to $\frac{7}{8}$, and k ranges from 3 to several hundred [Clark and Cain, 1981].
- *Convolutional codes.* A convolutional code is produced by a coder that has *memory.* The convolutional coder accepts k binary symbols at its input and produces n binary symbols at its output, where the n output symbols are affected by $v + k$ input symbols. Memory is incorporated because $v > 0$. The code rate is defined by $R = k/n$. Typical values for k and n range from 1 to 8, and the values for v range from 2 to 60. The range of R is between $\frac{1}{4}$ and $\frac{7}{8}$ [Clark and Cain, 1981]. A small value for the code rate R indicates a high degree of redundancy, which should provide more effective error control at the expense of increasing the bandwidth of the encoded signal.

Block Codes

Before discussing block codes, several definitions are needed. The *Hamming weight* of a code word is the number of binary 1 bits. For example, the code word 110101 has a Hamming weight of 4. The *Hamming distance* between two code words, denoted by d, is the number of positions by which they differ. For example, the code words 110101 and 111001 have a distance of $d = 2$. A received code word can be checked for errors. Some of the errors can be

† Do not confuse the code rate (with units of bits/bits) with the data rate or information rate (which has units of bits/s).

detected and corrected if $d \geq s + t + 1$, where s is the number of errors that can be detected and t is the number of errors that can be corrected ($s \geq t$). Thus, a pattern of t or fewer errors can be both detected and corrected if $d \geq 2t + 1$.

A general code word can be expressed in the form

$$i_1 i_2 i_3 \ldots i_k p_1 p_2 p_3 \ldots p_r$$

where k is the number of information bits, r is the number of parity check bits, and n is the total word length in the (n, k) block code, where $n = k + r$. This arrangement of the information bits at the beginning of the code word followed by the parity bits is most common. Such a block code is said to be *systematic*. Other arrangements with the parity bits interleaved between the information bits are possible and are usually considered to be equivalent codes.

Hamming has given a procedure for designing block codes that have single error-correction capability [Hamming, 1950]. A Hamming code is a block code having a Hamming distance of 3. Because $d \geq 2t + 1, t = 1$, and a single error can be detected and corrected. However, only certain (n, k) codes are allowable. These allowable Hamming codes are

$$(n, k) = (2^m - 1, 2^m - 1 - m) \tag{1–11}$$

where m is an integer and $m \geq 3$. Thus, some of the allowable codes are (7, 4), (15, 11), (31, 26), (63, 57), and (127, 120). The code rate R approaches 1 as m becomes large.

In addition to Hamming codes, there are many other types of block codes. One popular class consists of the cyclic codes. *Cyclic codes* are block codes, such that another code word can be obtained by taking any one code word, shifting the bits to the right, and placing the dropped-off bits on the left. These types of codes have the advantage of being easily encoded from the message source by the use of inexpensive linear shift registers with feedback. This structure also allows these codes to be easily decoded. Examples of cyclic and related codes are Bose–Chaudhuri–Hocquenhem (BCH), Reed-Solomon, Hamming, maximal–length, Reed–Müller, and Golay codes. Some properties of block codes are given in Table 1–3 [Bhargava, 1983].

TABLE 1–3 PROPERTIES OF BLOCK CODES

	Code[a]			
Property	BCH	Reed-Solomon	Hamming	Maximal Length
Block length	$n = 2^m = 1$ $m = 3, 4, 5, \ldots$	$n = m(2^m - 1)$ bits	$n = 2^m - 1$	$n = 2^m - 1$
Number of parity bits		$r = m2t$ bits	$r = m$	
Minimum distance	$d \geq 2t + 1$	$d = m(2t + 1)$ bits	$d = 3$	$d = 2^m - 1$
Number of information bits	$k \geq n - mt$			$k = m$

[a] m is any positive integer unless otherwise indicated; n is the block length; k is the number of information bits.

Convolutional Codes

A convolutional encoder is illustrated in Fig. 1–5. Here k bits (one input frame) are shifted in each time, and, concurrently, n bits (one output frame) are shifted out, where $n > k$. Thus, every k-bit input frame produces an n-bit output frame. Redundancy is provided in the output, because $n > k$. Also, there is memory in the coder, because the output frame depends on the previous K input frames, where $K > 1$. The *code rate* is $R = k/n$, which is $\frac{3}{4}$ in this illustration. The *constraint length*, K, is the number of input frames that are held in the kK-bit shift register.† Depending on the particular convolutional code that is to be generated, data from the kK stages of the shift register are added (modulo 2) and used to set the bits in the n-stage output register.

For example, consider the convolutional coder shown in Fig. 1–6. Here, $k = 1, n = 2, k = 3$, and a commutator with two inputs performs the function of a two-stage output shift register. The convolutional code is generated by inputting a bit of data and then giving the commutator a complete revolution. The process is repeated for successive input bits to produce the convolutionally encoded output. In this example, each $k = 1$ input bit produces $n = 2$ output bits, so the code rate is $R = k/n = \frac{1}{2}$. The code tree of Fig. 1–7 gives the encoded sequences for the convolutional encoder example of Fig. 1–6. To use the code tree, one moves up if the input is a binary 0 and down if the input is a binary 1. The corresponding encoded bits are shown in parentheses. For example, if

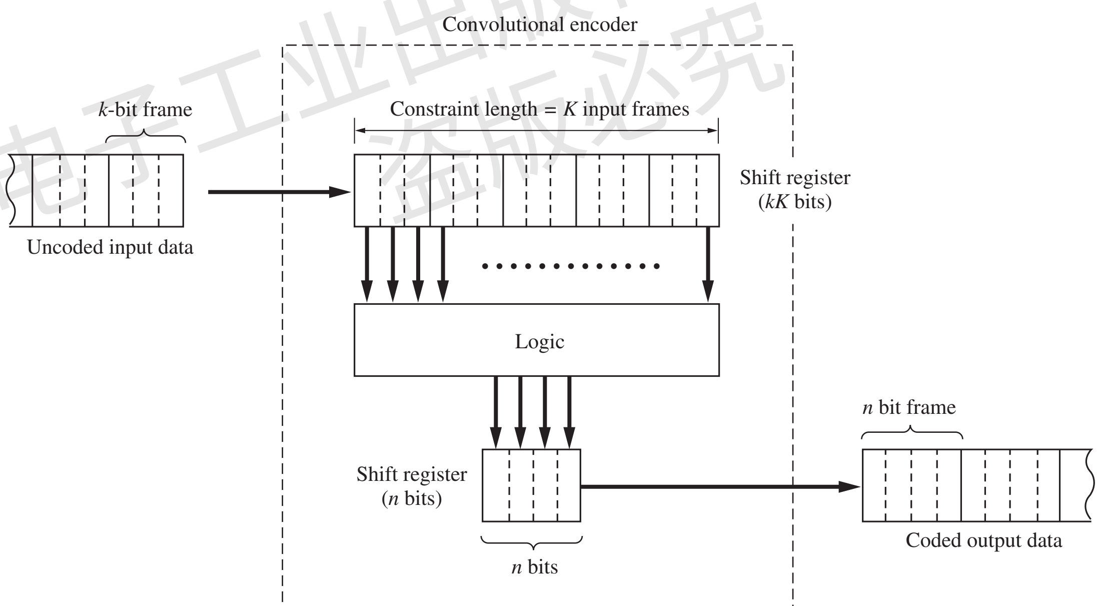


Figure 1–5 Convolutional encoding ($k = 3$, $n = 4$, $K = 5$, and $R = \frac{3}{4}$).

† Several different definitions of constraint length are used in the literature [Blahut, 1983; Clark and Cain, 1981; Proakis, 1995].

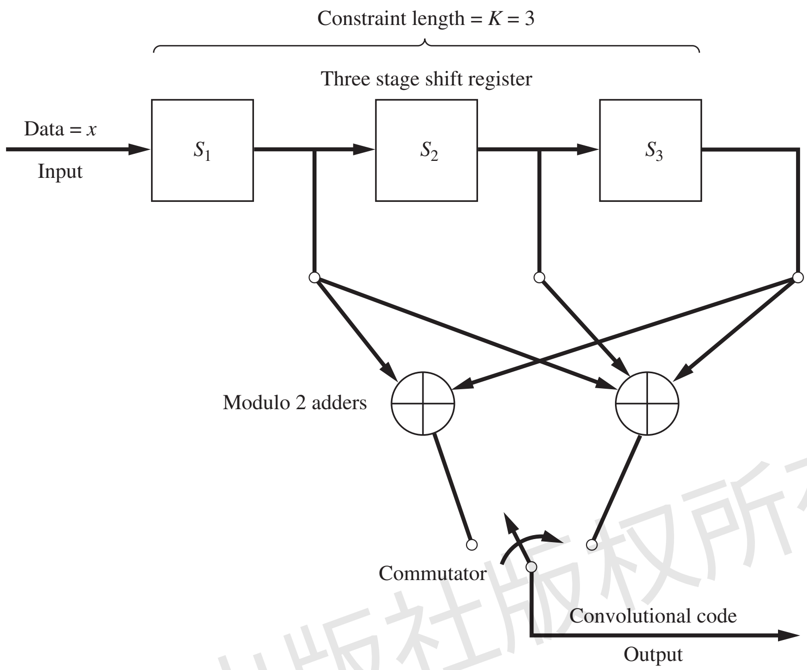


Figure 1–6 Convolutional encoder for a rate $\frac{1}{2}$, constraint length 3 code.

the input sequence $x_{11} = 1010$ is fed into the input (with the most recent input bit on the right), the corresponding encoded output sequence is $y_{11} = 11010001$, as shown by path A in Fig. 1–7.

A convolutionally encoded signal is decoded by "matching" the encoded received data to the corresponding bit pattern in the code tree. In sequential decoding (a suboptimal technique), the path is found like that of a driver who occasionally makes a wrong turn at a fork in a road but discovers the mistake, goes back, and tries another path. For example, if $y_{11} = 11010001$ was received, path A would be the closest match, and the decoded data would be $x_{11} = 1010$. If noise was present in the channel, some of the received encoded bits might be in error, and then the paths would not match exactly. In this case, the match is found by choosing a path that will minimize the Hamming distance between the selected path sequence and the received encoded sequence.

An optimum decoding algorithm, called *Viterbi decoding*, uses a similar procedure. It examines the possible paths and selects the best ones, based on some conditional probabilities [Forney, 1973]. The Viterbi procedure can use either soft or hard decisions. A *soft-decision* algorithm first decides the result on the basis of the test statistic[†] being above or below a decision threshold and then gives a "confidence" number that specifies how close the test statistic was to the threshold value. In *hard decisions*, only the

[†] The test statistic is a value that is computed at the receiver, based on the receiver input during some specified time interval. [See Eq. (7–4).]

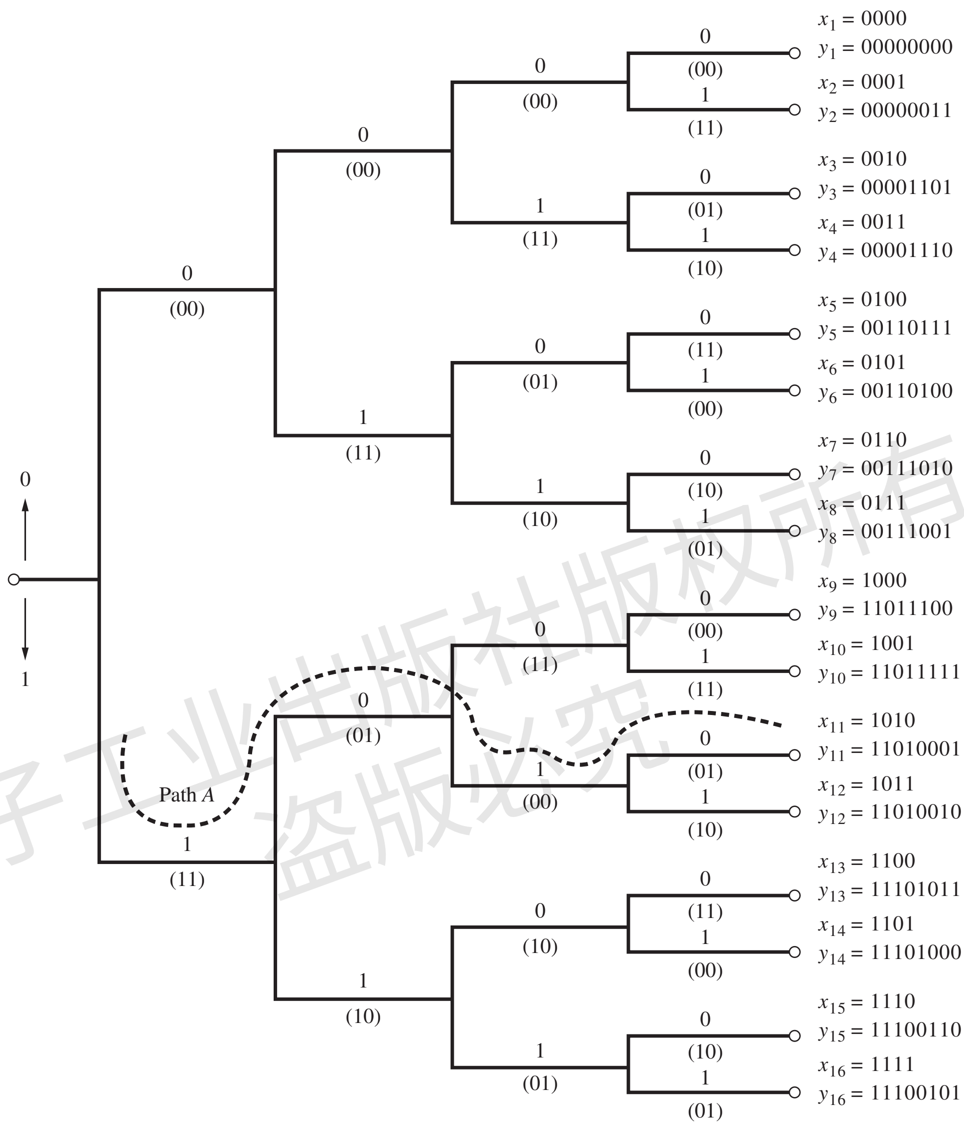


Figure 1–7 Code tree for convolutional encoder of Fig. 1–6.

decision output is known, and it is not known if the decision was almost "too close to call" (because the test value was almost equal to the threshold value). The soft-decision technique can translate into a 2-dB improvement (decrease) in the required receiver input E_b/N_0 [Clark and Cain, 1981]. E_b is the received signal energy over a 1-bit time interval, and $N_0/2$ is the power spectral density (PSD) of the channel noise at the receiver input. Both E_b and N_0 will be defined in detail in later chapters. [E.g. see Eq. (7–24b) or Eq. (8–44).]

Code Interleaving

In the previous discussion, it was assumed that if no coding was used, the channel noise would cause random bit errors at the receiver output that are more or less isolated (i.e., not adjacent). When coding was added, redundancy in the code allowed the receiver decoder to correct the errors so that the decoded output was almost error free. However, in some applications, large, wide pulses of channel noise occur. If the usual coding techniques are used in these situations, bursts of errors will occur at the decoder output because the noise bursts are wider than the "redundancy time" of the code. This situation can be ameliorated by the use of code interleaving.

At the transmitting end, the coded data are interleaved by shuffling (i.e., like shuffling a deck of cards) the coded bits over a time span of several block lengths (for block codes) or several constraint lengths (for convolutional codes). The required span length is several times the duration of the noise burst. At the receiver, before decoding, the data with error bursts are deinterleaved to produce coded data with isolated errors. The isolated errors are then corrected by passing the coded data through the decoder. This produces almost error-free output, even when noise bursts occur at the receiver input. There are two classes of interleavers—block interleavers and convolutional interleavers [Sklar, 1988].

Code Performance

The improvement in the performance of a digital communication system that can be achieved by the use of coding is illustrated in Fig. 1–8. It is assumed that a digital signal plus channel noise is present at the receiver input. The performance of a system that uses binary-phase-shift-keyed (BPSK) signaling is shown both for the case when coding is used and for the case when there is no coding. For the no-code case, the optimum (matched filter) detector circuit is used in the receiver, as derived in Chapter 7 and described by Eq. (7–38). For the coded case, a (23, 12) Golay code is used. P_e is the *probability of bit error*—also called the *bit error rate* (BER)—that is measured at the receiver output. E_b/N_0 is the energy-per-bit/noise-density ratio at the receiver input (as described in the preceding section). For $E_b/N_0 = 7\,\text{dB}$, Fig. 1–8 shows that the BER is 10^{-3} for the uncoded case and that the BER can be reduced to 10^{-5} if coding is used.

The *coding gain* is defined as the reduction in E_b/N_0 (in decibels) that is achieved when coding is used, when compared with the E_b/N_0 required for the uncoded case at some specific level of P_e. For example, as can be seen in the figure, a coding gain of 1.33 dB is realized for a BER of 10^{-3}. The coding gain increases if the BER is smaller, so that a coding gain of 2.15 dB is achieved when $P_e = 10^{-5}$. This improvement is significant in space communication applications, where every decibel of improvement is valuable. The figure also shows noted that there is a *coding threshold* in the sense that the coded system actually provides *poorer* performance than the uncoded system when E_b/N_0 is less than the threshold value. In this example, the coding threshold is about 3.5 dB. A coding threshold is found in all coded systems.

For optimum coding, Shannon's channel capacity theorem, Eq. (1–10), gives the E_b/N_0 required. That is, if the source rate is below the channel capacity, the optimum code will allow the source information to be decoded at the receiver with $P_e \to 0$ (i.e., $10^{-\infty}$), even though

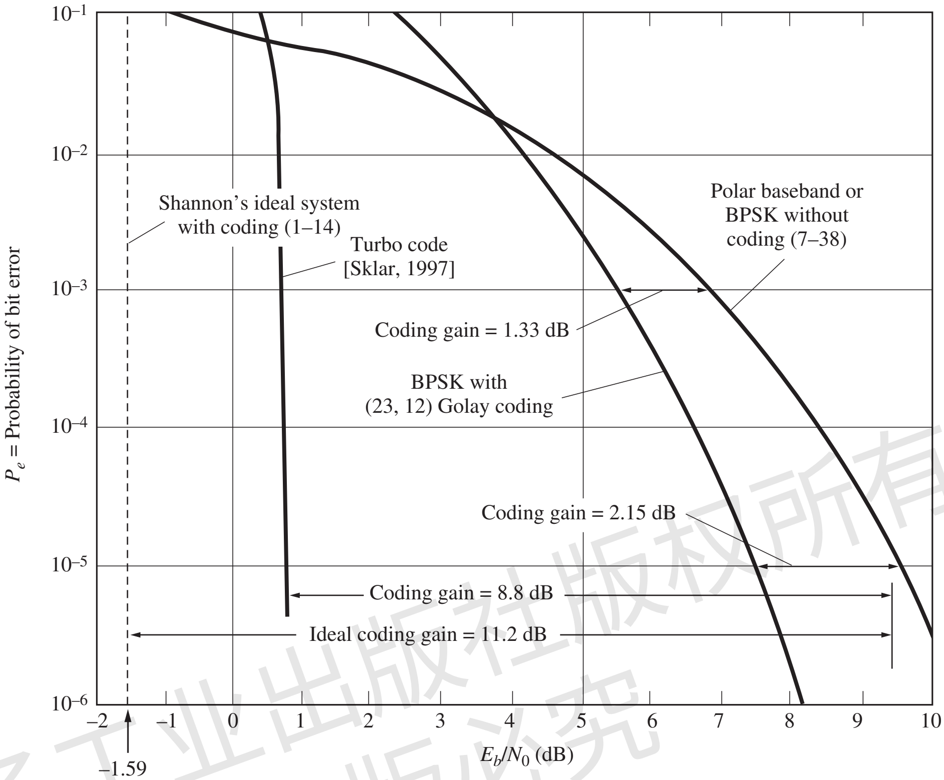


Figure 1–8 Performance of digital systems—with and without coding.

there is some noise in the channel. We will now find the E_b/N_0 required so that $P_e \to 0$ with the optimum (unknown) code. Assume that the optimum encoded signal is not restricted in bandwidth. Then, from Eq. (1–10),

$$C = \lim_{B\to\infty}\left\{B\log_2\left(1+\frac{S}{N}\right)\right\} = \lim_{B\to\infty}\left\{B\log_2\left(1+\frac{E_b/T_b}{N_0B}\right)\right\}$$

$$= \lim_{x\to 0}\left\{\frac{\log_2\left[1+(E_b/N_0T_b)x\right]}{x}\right\}$$

where T_b is the time that it takes to send one bit and N is the noise power that occurs within the bandwidth of the signal. The power spectral density (PSD) is $\mathcal{P}_n(f) = N_0/2$, and, as shown in Chapter 2, the noise power is

$$N = \int_{-B}^{B}\mathcal{P}_n(f)\,df = \int_{-B}^{B}\left(\frac{N_0}{2}\right)df = N_0B \tag{1–12}$$

where B is the signal bandwidth. L'Hospital's rule is used to evaluate this limit:

$$C = \lim_{x \to 0} \left\{ \frac{1}{1 + (E_b/N_0 T_b)x} \left(\frac{E_b}{N_0 T_b} \right) \log_2 e \right\} = \frac{E_b}{N_0 T_b \ln 2} \tag{1–13}$$

If we signal at a rate approaching the channel capacity, then $P_e \to 0$, and we have the maximum information rate allowed for $P_e \to 0$ (i.e., the optimum system). Thus, $1/T_b = C$, or, using Eq. (1–13),

$$\frac{1}{T_b} = \frac{E_b}{N_0 T_b \ln 2}$$

or

$$E_b/N_0 = \ln 2 = -1.59\,\text{dB} \tag{1–14}$$

This minimum value for E_b/N_0 is -1.59 dB and is called Shannon's limit. That is, if optimum coding/decoding is used at the transmitter and receiver, error-free data will be recovered at the receiver output, provided that the E_b/N_0 at the receiver input is larger than -1.59 dB. This "brick wall" limit is shown by the dashed line in Fig. 1–8, where *Pe* jumps from 0 ($10^{-\infty}$) to $\frac{1}{2}(0.5 \times 10^0)$ as E_b/N_0 becomes smaller than -1.59 dB, assuming that the ideal (unknown) code is used. Any practical system will perform worse than this ideal system described by Shannon's limit. Thus, the goal of digital system designers is to find practical codes that approach the performance of Shannon's ideal (unknown) code.

When the performance of the optimum encoded signal is compared with that of BPSK without coding (10^{-5} BER), it is seen that the optimum (unknown) coded signal has a coding gain of $9.61 - (-1.59) = 11.2$ dB. Using Fig. 1–8, compare this value with the coding gain of 8.8 dB that is achieved when a turbo code is used. Table 1–4 shows the gains that can be obtained for some other codes.

Since their introduction in 1993, *turbo codes* have become very popular because they can perform near Shannon's limit, yet they also can have reasonable decoding complexity [Sklar, 1997]. Turbo codes are generated by using the parallel concatenation of two simple convolutional codes, with one coder preceded by an interleaver [Benedetto and Montorsi, 1996]. The interleaver ensures that error-prone words received for one of the codes corresponds to error-resistant words received for the other code.

All of the codes described earlier achieve their coding gains at the expense of *bandwidth expansion.* That is, when redundant bits are added to provide coding gain, the overall data rate and, consequently, the bandwidth of the signal are increased by a multiplicative factor that is the reciprocal of the code rate; the bandwidth expansion of the coded system relative to the uncoded system is $1/R = n/k$. Thus, if the uncoded signal takes up all of the available bandwidth, coding cannot be added to reduce receiver errors, because the coded signal would take up too much bandwidth. However, this problem can be ameliorated by using trellis-coded modulation (TCM).

Trellis-Coded Modulation

Gottfried Ungerboeck has invented a technique called *trellis-coded modulation* (TCM) that combines multilevel modulation with coding to achieve coding gain without bandwidth

TABLE 1–4 CODING GAINS WITH POLAR BASEBAND, BPSK OR QPSK

Coding Technique Used	Coding Gain (dB) at 10^{-5} BER	Coding Gain (dB) at 10^{-8} BER	Data Rate Capability
Ideal coding	11.2	13.6	
Turbo code [Sklar, 1997]	8.8		
Concatenated[a] Reed–Solomon and convolution (Viterbi decoding)	6.5–7.5	8.5–9.5	Moderate
Convolution with sequential decoding (soft decisions)	6.0–7.0	8.0–9.0	Moderate
Block codes (soft decision)	5.0–6.0	6.5–7.5	Moderate
Concatenated[a] Reed–Solomon and short block	4.5–5.5	6.5–7.5	Very high
Convolutional with Viterbi decoding	4.0–5.5	5.0–6.5	High
Convolutional with sequential decoding (hard decisions)	4.0–5.0	6.0–7.0	High
Block codes (hard decisions)	3.0–4.0	4.5–5.5	High
Block codes with threshold decoding	2.0–4.0	3.5–5.5	High
Convolutional with threshold decoding	1.5–3.0	2.5–4.0	Very high

[a]Two different encoders are used in series at the transmitter (see Fig. 1–4), and the corresponding decoders are used at the receiver.

Source: Bhargava [1983] and Sklar [1997].

expansion [Benedetto, Mondin, and Montorsi, 1994; Biglieri, Divsalar, McLane, and Simon, 1991; Ungerboeck, 1982, 1987]. The trick is to add the redundant coding bits by increasing the number of levels (amplitude values) allowed in the digital signal without changing the pulse width. (The bandwidth will remain the same if the pulse width is not changed, since the bandwidth is proportional to the reciprocal of the pulse width.) This technique is called multilevel signaling and is first introduced in Section 3–4. For example, the pulses shown in Fig. 3–14a represent $L = 4$ multilevel signaling, where each level carries two bits of information as shown in Table 3–3. Now add one redundant coding bit to the two information bits to provide eight amplitude levels for the pulses, but maintaining the same pulse width so that the waveform would have the same bandwidth. Then the redundant bit, due to the coding, could be accommodated without any increase in bandwidth. This concept can be generalized to complex-valued multilevel signaling, as shown at the end of Sec. 5–10. In summary, TCM integrates waveform modulation design, with coding design, while maintaining the bandwidth of the uncoded waveform.

When a convolutional code of constraint length $K = 3$ is implemented, this TCM technique produces a coding gain of 3 dB relative to an uncoded signal that has the same bandwidth and information rate. Almost 6 dB of coding gain can be realized if coders of constraint length 9 are used. The larger constraint length codes are not too difficult to generate, but the corresponding decoder for a code of large constraint length is very complicated. However, very-high-speed integrated circuits (VHSIC) make this such a decoder feasible.

The 9,600-bit/s CCITT V.32, 14,400-bit/s CCITT V.33bis, and 28,800-bit/s CCITT V.34 computer modems use TCM. The CCITT V.32 modem has a coding gain of 4 dB and is described by Example 4 of Wei's paper [Wei, 1984; CCITT Study Group XVII, 1984].

For further study about coding, the reader is referred to several excellent books on the topic [Blahut, 1983; Clark and Cain, 1981; Gallagher, 1968; Lin and Costello, 1983; McEliece, 1977; Peterson and Weldon, 1972; Sweeney, 1991; Viterbi and Omura, 1979].

1–12 PREVIEW

From the previous discussions, we see the need for some basic tools to understand and design communication systems. Some prime tools that are required are mathematical models to represent signals, noise, and linear systems. Chapter 2 provides these tools. It is divided into the broad categories of properties of signal and noise, Fourier transforms and spectra, orthogonal representations, bandlimited representations, and descriptions of linear systems. Measures of bandwidth are also defined.

1–13 STUDY-AID EXAMPLES

SA1–1 Evalution of Line of Sight (LOS) The antenna for a television (TV) station is located at the top of a 1,500-foot transmission tower. Compute the LOS coverage for the TV station if the receiving antenna (in the fringe area) is 20 feet above ground.

Solution: Using Eq. (1–6), we find that the distance from the TV transmission tower to the radio horizon is

$$d_1 = \sqrt{2h} = \sqrt{2(1{,}500)} = 54.8\,\text{miles}$$

The distance from the receiving antenna to the radio horizon is

$$d_2 = \sqrt{2(20)} = 6.3\,\text{miles}$$

Then, the total radius for the LOS coverage contour (which is a circle around the transmission tower) is

$$d = d_1 + d_2 = 61.1\,\text{miles}$$

SA1–2 Information Data Rate A telephone touch-tone keypad has the digits 0 to 9, plus the * and # keys. Assume that the probability of sending * or # is 0.005 and the probability of sending 0 to 9 is 0.099 each. If the keys are pressed at a rate of 2 keys/s, compute the data rate for this source.

Solution: Using Eq. (1–8), we obtain

$$H = \Sigma P_j \log_2 \left(\frac{1}{P_j} \right)$$

$$= \frac{1}{\log_{10}(2)} \left[10(0.099) \log_{10} \left(\frac{1}{0.099} \right) + 2(0.005) \log_{10} \left(\frac{1}{0.005} \right) \right]$$

or

$$H = 3.38 \text{ bits/key}$$

Using Eq. (1–9), where $T = 1/(2 \text{ keys/s}) = 0.5 \text{ s/key}$, yields

$$R = \frac{H}{T} = \frac{3.38}{0.5} = 6.76 \text{ bits/s}$$

SA1–3 Maximum Telephone Line Data Rate A computer user plans to buy a higher-speed modem for sending data over his or her analog telephone line. The telephone line has a signal-to-noise ratio (SNR) of 25 dB and passes audio frequencies over the range from 300 to 3,200 Hz. Calculate the maximum data rate that could be sent over the telephone line when there are no errors at the receiving end.

Solution: In terms of a power ratio, the SNR is $S/N = 10^{(25/10)} = 316.2$ (see dB in Chapter 2), and the bandwidth is $B = 3{,}200 - 300 = 2{,}900$ Hz. Using Eq. (1–10), we get

$$R = B \log_2 \left(1 + \frac{S}{N} \right) = 2{,}900 \, [\log_{10}(1 + 316.2)]/\log_{10}(2),$$

or

$$R = 24{,}097 \text{ bits/s}$$

Consequently, a 28.8-kbit/s modem signal would not work on this telephone line; however, a 14.4-kbit/s modem signal should transmit data without error.

PROBLEMS

1–1 Assume that the Earth's terrain is relatively flat but that there is a mountain that has a height of 800 feet above the average terrain. How far away could a 100-ft cell tower be located from this mountain and yet provide cell phone coverage to a person on top of this mountain?

1–2 A high-power FM station of frequency 96.9 MHz has an antenna height of 1000 ft. If the signal is to be received 55 miles from the station, how high does a prospective listener need to mount his or her antenna in this fringe area?

★ **1–3** Using geometry, prove that Eq. (1–6) is correct.

1–4 A terrestrial microwave system is being designed. The transmitting and receiving antennas are to be placed at the top of equal-height towers, with one tower at the transmitting site and one at the receiving site. The distance between the transmitting and receiving sites is 25 miles. Calculate the minimum tower height required for an LOS transmission path.

1–5 A cellular telephone cell site has an antenna located at the top of a 100-ft tower. A typical cellular telephone user has his or her antenna located 4 ft above the ground. What is the LOS radius of coverage for this cell site to a distant user?

★ **1–6** A digital source emits -1.0-and 0.0-V levels with a probability of 0.2 each and $+3.0$- and $+4.0$-V levels with a probability of 0.3 each. Evaluate the average information of the source.

1–7 Prove that base 10 logarithms may be converted to base 2 logarithms by using the identity $\log_2(x) = [1/\log_{10}(2)]\log_{10}(x)$.

1–8 If all the messages emitted by a source are equally likely (i.e., $P_j = P$), show that Eq. (1–8) reduces to $H = \log_2(1/P)$.

★ **1–9** For a binary source:

(a) Show that the entropy H is a maximum when the probability of sending a binary 1 is equal to the probability of sending a binary 0.

(b) Find the value of maximum entropy.

★ **1–10** A single-digit, seven-segment liquid crystal display (LCD) emits a 0 with a probability of 0.25; a 1 and a 2 with a probability of 0.15 each; 3, 4, 5, 6, 7, and 8 with a probability of 0.07 each; and a 9 with a probability of 0.03. Find the average information for this source.

1–11 **(a)** A binary source sends a binary 1 with a probability of 0.3. Evaluate the average information for the source.

(b) For a binary source, find the probability for sending a binary 1 and a binary 0, such that the average source information will be maximized.

★ **1–12** A numerical keypad has the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9. Assume that the probability of sending any one digit is the same as that for sending any of the other digits. Calculate how often the buttons must be pressed in order to send out information at the rate of 3 bits/s.

1–13 Refer to Example 1–3 and assume that words, each 12 digits in length, are sent over a system and that each digit can take on one of two possible values. Half of the possible words have a probability of being transmitted that is $(\frac{1}{2})^{13}$for each word. The other half have probabilities equal to $3(\frac{1}{2})^{13}$ Find the entropy for this source.

1–14 Evaluate the channel capacity for a teleprinter channel that has a 300-Hz bandwidth and an SNR of 30 dB.

★ **1–15** Assume that a computer terminal has 110 characters (on its keyboard) and that each character is sent by using binary words.

(a) What are the number of bits needed to represent each character?

(b) How fast can the characters be sent (characters/s) over a telephone line channel having a bandwidth of 3.2 kHz and an SNR of 20 dB?

(c) What is the information content of each character if each is equally likely to be sent?

1–16 An analog telephone line has an SNR of 45 dB and passes audio frequencies over the range of 300 to 3,200 Hz. A modem is to be designed to transmit and receive data simultaneously (i.e., full duplex) over this line without errors.

(a) If the frequency range 300 to 1,200 Hz is used for the transmitted signal, what is the maximum transmitted data rate?

(b) If the frequency range 1,500 to 3,200 Hz is used for the signal being simultaneously received, what is the maximum received data rate?

(c) If the whole frequency range of 300 to 3,200 Hz is used simultaneously for transmitting and receiving (by the use of a hybrid circuit as described in Chapter 8, Fig. 8–4), what are the maximum transmitting and receiving data rates?

1–17 Using the definitions for terms associated with convolutional coding, draw a block diagram for a convolutional coder that has rate $R = \frac{2}{3}$ and constraint length $K = 3$.

★ **1–18** For the convolutional encoder shown in Fig. P1–18, compute the output coded data when the input data is $x = [10111]$. (The first input bit is the leftmost element of the **x** row vector.)

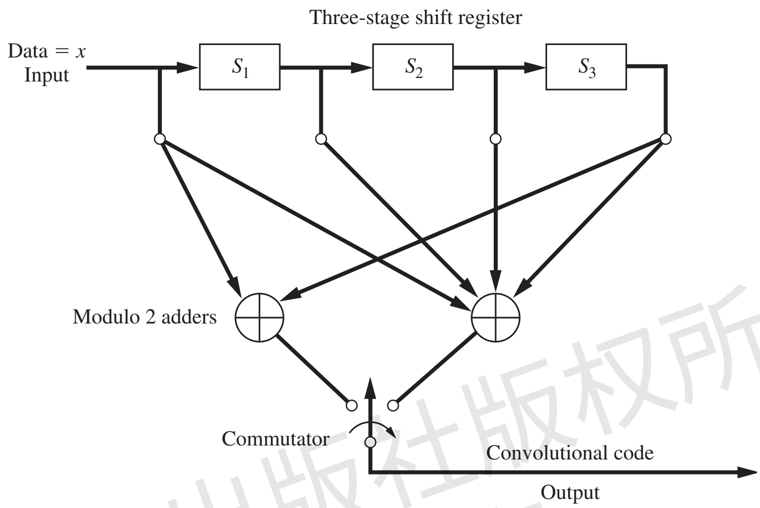


Figure P1–18