1

An Overview

Modern society obviously depends on electronic communication for much of its functioning. Among the many possible ways of communicating, the class of techniques referred to as ***digital communications*** has become predominant in the latter part of the 20th century, and indications are that this trend will continue. There are a number of important conceptual reasons for this development, as well as some related to the advance of technology and to economics, and we will discuss these shortly. First, however, we should gain a working understanding of the digital communications process.

Digital communication is simply the practice of exchanging information by the use of ***finite sets of signals***. In modern practice these signals are in the form of electrical waveforms or electromagnetic fields. The essence of digital communication can easily be captured, however, by recalling more primitive "digital" techniques, say smoke signaling or the use of colored flags in the nautical world. The allowable set of signals, represented by patterns of puffs of smoke or by flag color and position, is finite, and the users are communicating in a digital manner, albeit with a slow, rather unreliable method. As a side note, the word ***digital*** often suggests the number 10, but this usage is too restrictive; perhaps the term ***discrete communications*** is more apt.

Nondigital means of communication are known as analog communication and have historically been more prevalent, although the first important electrical communication device, the telegraph, is another simple digital example. A familiar analog system is the traditional telephone network. The speech signal originates as an acoustic pressure wave in our vocal tract, and an electrical analog of this pressure wave, produced by a microphone, is sent through the switched telephone network. Another illustration is television (seeing at a distance), where the intensity of a radio signal is varied, or modulated, by the output of a scanning camera. In both situations, the message to be communicated is presumably one from an infinite collection. Although the visual or auditory system is incapable of resolving an arbitrarily large number of signals in a given interval of time, the analog communication process proceeds as if such were possible.[1]

Another distinguishing trait is that analog systems are for the most part operating in continuous time, whereas digital schemes always involve events on a discrete time line. That is, we agree to send a new signal from our finite set every T_s seconds, where T_s is called the signaling interval. The signaling interval is application-dependent, typically ranging between milliseconds and nanoseconds. Surprisingly perhaps, this discretization of both time and signal sets costs us nothing in terms of ultimate communication efficiency, a fundamental result of information theory. We will return to this notion shortly.

The increasing popularity of digital telecommunication techniques is due to several factors.* The most prominent are as follows:

1. The use of digital transmission supports the electronic ***addressing and routing*** of messages in a multiuser system, for example in distributed electronic mail networks.

2. Digital messages associated with speech, video, or alphanumeric data files appear essentially identical, although their data rates may vary. Therefore, different forms of information are easily accommodated by a digital transmission facility, or many separate and disparate sources can be multiplexed into one aggregate digital message. This provides increased ***flexibility***, or ***multimedia capability***. An illustration is digital telephony, where N channels of audio ($N = 24$ and 32 in North America/Japan and Europe, respectively) are multiplexed into a single bit stream having a transmission rate of 1.544 (or 2.048) megabits per second, respectively. These bit streams may in turn be combined with other similar signals at higher levels in a multiplexing hierarchy, or, alternatively, a digitized color television signal could perhaps substitute for N channels of audio. Still another option would be to substitute a computer-to-computer connection for an audio channel.

3. Digital messages are more easily ***encrypted*** than analog waveforms. Such encryption can have two purposes: to make the message unreadable by unintended recipients and to provide electronic authentication of the sending party.

4. Digital messages may be accurately and rapidly ***stored and retrieved*** electronically, whereas in the analog realm, we are faced with tape recorders, photographic film, and the like, which are beset with slow access time and lesser data integrity.

5. In progressing through a transmission system with several hops, or stages, the digital message may be ***reconstituted*** at each stage, in contrast to the progressive accumu-

[1] Actually, the telephone system is steadily becoming more digital in nature worldwide, and high-definition television will likely involve many of the digital transmission principles discussed here.

lation of noise and distortion in traditional analog systems, such as multihop microwave FM telephony systems. The TAT-8 transatlantic fiber-optic system employs some 100 repeaters at the bottom of the ocean to regenerate en route the high-bit-rate digital optical signal, which conveys some 40,000 simultaneous conversations.

Incidentally, the usual homage to digital technology, something like "digital circuits are more reliable and less expensive than analog circuits designed to do the same thing," is not germane here, although one would probably concede its truth. We are not concerned here with the circuit-level implementation of systems; in fact, a large part of the important signal processing in digital communication systems must be done with analog components or by making digital approximations to analog operations.

With some knowledge of the "what" and "why" aspects of digital communications, let's now begin to see "how" it's accomplished.

1.1 A FRAMEWORK FOR DIGITAL COMMUNICATIONS

In this text we shall address issues of ***single-source/single-destination*** digital communication, although much of the practical interest in this material derives from multiuser applications building upon our treatment. A generic model for such a point-to-point digital communication system is shown in Figure 1.1.1. The givens of the system are the ***information source***, or message generator; the ***channel***, or physical medium by which communication is to take place; and the ***user***, or information sink. These system elements are emphasized in shaded boxes, and are presumed to be the parts of the system over which we have no control. We shall say more shortly about the other elements of Figure 1.1.1, on which we can exert considerable design influence.

1.1.1 Sources, Channels, and Limits to Communication

The ***source*** may inherently be a discrete (or digital) source, such as an alphanumeric keyboard generating a message, or it may produce a sequence of real-valued samples as its message. In either case, elements of the source output sequence will be designated W_n. A third possibility, often the case in practice, is that the source output is an electrical waveform $W(t)$, continuous in amplitude and time, as in the example of a speech signal produced by a microphone. In any situation, however, the information source is modeled probabilistically, and we will view messages as outputs from some random experiment. (If messages to be sent were produced by a completely deterministic process, there is in fact no information to be conveyed! Note this does not imply that human beings have no causality or intent behind what they say or write; to potential recipients, however, there is simply a priori uncertainty about the message to be received.)

The ***channel*** should be broadly understood as a physical mechanism that accepts an input signal, denoted $S(t)$ in Figure 1.1.1, and produces an output signal, $R(t)$, which in general is an imperfect rendition of $S(t)$. Our waveform-level view of the channel attempts to address the true processes of the channel, although popular discrete-time, discrete-alphabet models for channels can be derived from the waveform counterparts.

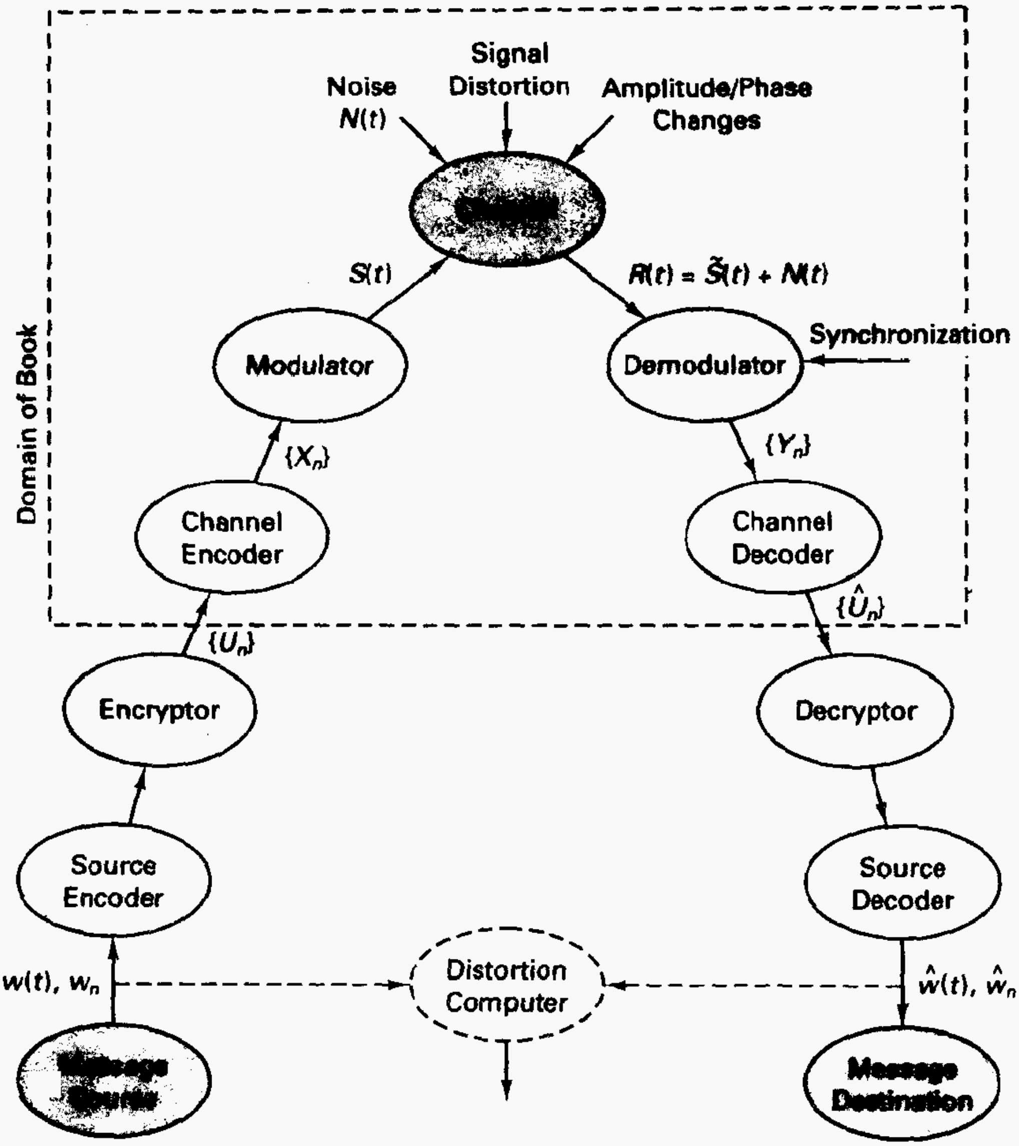


Figure 1.1.1 Digital communication pathway.

The corruption of the signal is typically of two forms:

1. The addition of ***noise*** by electronic equipment used to perform the communication process, or the influence of external noise processes such as cosmic and atmospheric electromagnetic noise or interfering signals,
2. Channel ***distortions*** due to physical channel limitations (e.g., the bandwidth limitations of the voiceband telephone channel, or a magnetic tape recorder/player), or due to communication equipment again, such as filters or amplifiers.

In any case, we assume there is a well-defined mathematical model, which includes deterministic and stochastic aspects, for the action of the channel on the input signal.

There is usually some ambiguity over what constitutes the channel and what is properly part of the other boxes of Figure 1.1.1, that is, the demodulator, decoder, and others. A perfectly reasonable operational understanding was once made by J. L. Kelly—the channel is that part of the transmission system that we cannot change, or don't wish to change. As illustration, consider a digital tape recording system as a communication medium. (Here communication is not to some distant point in near real time, but to perhaps the same point in space at a later time, and such channels are called storage channels.) We might define the channel to include only the magnetics, that is, the read/write heads and the magnetic oxide passing near them. In this view the channel occupies a rather small volume of the entire recorder. On the other hand, if we purchase a tape recorder and wish to archive digital data with the recorder "as is," we would define the channel to be everything between the input/output connectors. The difference in definitions includes substantial electronic processing.

Another example raising similar ambiguity concerning channel specification involves optical signaling with a laser. Our best possibilities for efficient design remain when we process the electromagnetic signal directly; however, current technology typically allows that we observe the output of a photodetector, which converts optical photons into electrical current. Such detectors themselves are invariably noisy, exhibiting signal-dependent shot noise and dark currents and often have distorting effects upon the transmitted signals due to response-time limitations. Here then the channel definition could be limited to the electromagnetic medium (perhaps fiber-optic waveguide) or could incorporate a laser diode transmitter and a photodetector as well.

The ***user***, or destination, is relatively unnoteworthy, except that to the source/user tandem we may attach a fidelity criterion that describes goodness of performance. In analog systems, the criterion might be mean-square error between source and destination waveforms, but in discrete alphabet communications, the performance is more traditionally measured by quantities such as symbol error probability or message error probability.

Communication system design gets played out in one of two ways:

1. We are provided a channel with certain capabilities and wish to design a system that can provide communications at the largest aggregate rate, subject to tolerable distortion constraints.
2. The traffic load and required fidelity are specified, and we must engineer an efficient channel to accomplish this task. This normally involves designing transmitters, antennas, and receivers to supply a certain signal-to-noise ratio and bandwidth and, clearly, we would like to operate with minimal resources required to perform the job.

One of the supreme accomplishments of information theory, pioneered by Claude Shannon[2] in 1948 [1], is that armed only with a mathematical description of the source and channel, along with a measure of goodness, or fidelity criterion, relating source outputs $W(t)$ to user inputs or source reproductions, $\widehat{W}(t)$, (respectively W_n and $\widehat{W}_n$), we may determine the performance of the ultimate communication system (analog or

[2] The reader is strongly encouraged to study Shannon's two-part paper for historical perspective and for a lucid presentation of the mathematics of information transfer.

digital) without even describing its details! More precisely, every reasonable channel has a parameter, called the ***channel capacity***, C, measured in units of bits/unit time, which depends on the mathematical description of the channel, but in more familiar terms upon parameters such as signal-to-noise power ratio and upon available bandwidth. Similarly, for every combination of source model and fidelity criterion we can assign a ***rate distortion function*** $R(d)$, specified in bits per unit of time, that depends only on the source description and on the fidelity criterion. The argument, d, of the rate distortion function $R(d)$ is the smallest expected, or average, distortion achievable by any system representing the source with $R(d)$ bits per unit source time. The important connection occurs when we equate these two descriptors and obtain the solution for d, called d^*:

$$R(d^*) = C. \tag{1.1.1}$$

Information theory establishes the following two-edged result:

1. No system, no matter how complicated, can have an average distortion of less than d^*.
2. Systems may be built, if we allow suitably large delay and complexity in the communication process, that achieve average distortion arbitrarily close to d^*.

[If the d^* resulting from (1.1.1) is unacceptably large, then we must either reengineer the system to provide greater channel capacity or slow the source symbol production rate.]

In some sense, operating near this solution point is the holy grail of communication engineers, and attaining efficiency near this promise requires the communication engineer to be familiar with many different scenarios for formatting information suitable for different channels and with optimal means of recovering the message signal. It must be said, however, that for all but simple models of sources and channels the underlying mathematics needed to even formulate (1.1.1) is often intractable, and simple models often do not reflect reality too well. Nonetheless, the development of this result has provided us with the essence of good communication system design, that is, how to configure the remaining elements of Figure 1.1.1.

Another remarkable result of information theory is that, even when the source is an analog waveform or the channel is likewise analog, such as a microwave radio propagation medium with amplitude fading and receiver noise, no loss of optimality is incurred by adopting (proper) digital processing at the intervening levels. By this we do not mean simply that digital arithmetic (as in a computer) is allowed; we know this is allowable with suitable precision in calculation. Rather, the entire communication process can be effectively viewed as ***discrete alphabet, discrete time***, and thus we enter the arena of digital communications. In essence, nature forces us to accept a certain finiteness in our communication anyway, due to noise, sensory imperfections, power limitations, and the likes, and we may as well accept this a priori and do efficient signal processing with discrete message sets. thereby accruing the other operational benefits outlined previously.

1.1.2 Operations in the Digital Transmission Pathway

We now examine the role of the remaining modules in Figure 1.1.1. The ***source encoder's*** task is commonly referred to as data compression, although this description is open to misinterpretation, especially when the original signal is an analog waveform. Basically, the source encoder accepts the source outputs (whether a discrete sequence, real numbers, or waveforms) and produces a sequence of symbols, U_n, usually a sequence of bits,[3] that represents the source output in the best possible way (with respect to some distortion criterion) under the constraint of, say, allowing R_b bits per source unit of time. If the original source is discrete, source strings can often be perfectly represented (referred to as coded in noiseless or lossless fashion) by a more compact sequence of bits because of natural redundancies; hence the older term referring to this process is redundancy reduction. Modern facsimile transmission techniques for transmitting pages of text or graphical material exploit this redundancy between contiguous picture elements of the scanning process. At a higher level of representation, the text on this page may be more efficiently represented than by letter-at-a-time coding due to the relatively high frequency of certain digrams, trigrams, and even words such as "the" and "digital." Many modern data communication systems apply file compression algorithms to save transmission and/or storage requirements.

If the source is a sequence of real numbers, then some process of discretizing, or quantizing, is necessary prior to digital transmission. Finally, in the situation where the source produces a waveform, a sampling process to initially convert the signal to discrete time is conventional, although not formally necessary as an intermediate step. In many cases the source coding process is a many-to-one classification problem; that is, we map a large class of source messages onto a smaller discrete set of eventual source reproductions. In such cases, source encoding engenders eventual error in reconstructing the message sequence, although it is at least a controlled form of error or noise.

The ***source decoder*** performs a much simpler inverse process, essentially amounting to table lookup. It receives an identification string chosen by the source encoder (assuming no transmission errors) and outputs a message in appropriate form for the user (discrete alphabet characters, real numbers, or waveforms). Traditionally, this involves some form of digital-to-analog conversion and perhaps sample interpolation to produce waveforms.

In a formal sense, the source coding problem can be effectively decoupled from the other operations, notably channel modulation and coding. Fundamental results of information theory show that an efficient communication system, in the sense of (1.1.1), can be realized as a cascade of the following:

1. An efficient source encoder/decoder, which associates the source output with a discrete message set of source approximations, typically labeled by binary strings
2. An efficient channel modulation and coding system designed to convey these source coder labels

Our interest here is primarily in the latter (discrete source coding is discussed briefly in Chapter 2), and the point of view is that the source encoding task, if necessary, has been performed. Our primary intent is to design the system so that the sequence at the input to the channel encoder is correctly reproduced, with high probability, at the

[3] The use of *bit* for *binary digit* apparently first appears in Shannon's 1948 paper; Shannon credits J. W. Tukey for suggesting this contraction.

output of the channel decoder so that the end-to-end message distortion is normally dominated by the source encoding/decoding operations. This partitioning of the overall task is partly for conceptual reasons, and it may be that simple, integrated source/channel coding approaches work as well as those designed with heroic effort under the "separation principle." However, the separation is convenient in practical terms, for it makes explicit the ability of digital transmission facilities to handle a wide variety of sources, once converted into digital format. (Incidentally, despite many dualities between source and channel coding, practitioners seem to settle into one camp or the other!) Readers interested in the theoretical foundations of source coding, called rate distortion theory, are referred to Berger's text [2]. A treatment of source coding with a more applied orientation is found in the work of Jayant and Noll [3].

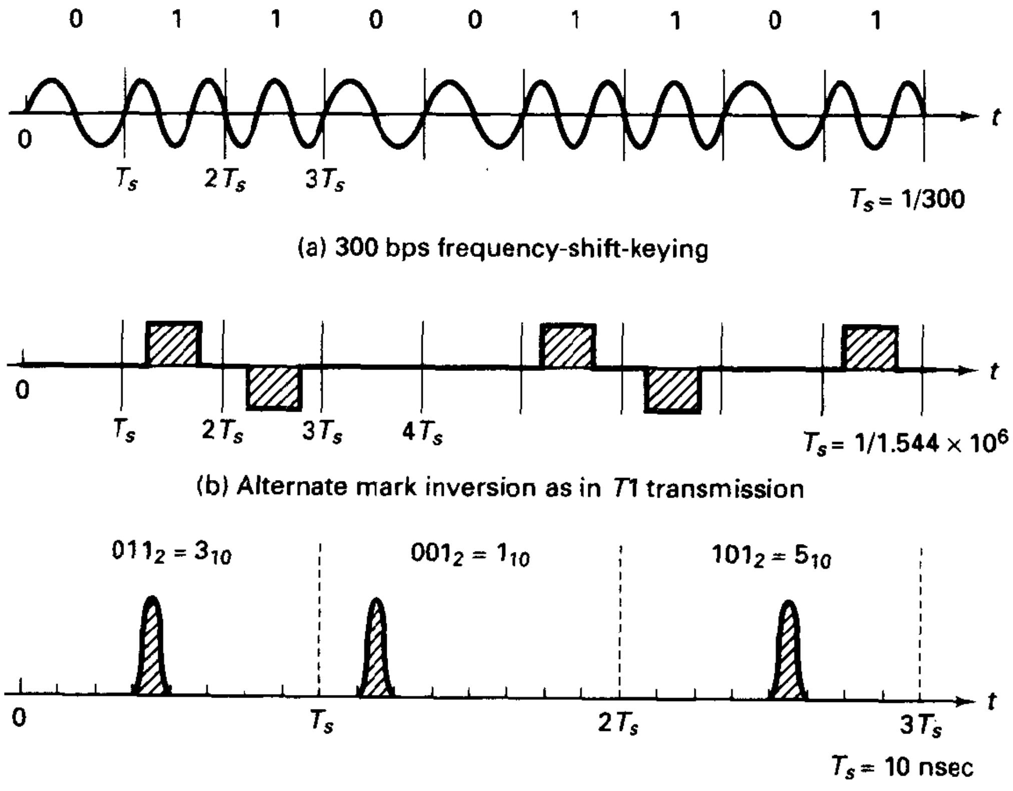
Another topic we shall not address here in detail is ***cryptography***, or secrecy coding. Encryption and decryption techniques were formerly relevant only to military and strategic governmental communications, but have lately gained importance in most aspects of telecommunications. Encryption devices map digital sequences of "plaintext" into "ciphertext" with the intent that reconstruction of the message without the key to these mappings is prohibitively difficult. (Decryption with the secret key is, however, easy.) Newer cryptographic techniques, called public-key systems, avoid the need for secure key distribution and can, in addition, provide an authentication or electronic signature feature, something digital transmission normally sacrifices! Konheim's text [4] provides a modern survey of this field.

1.1.3 Modulation and Coding

We then come to the ***channel encoder*** and ***modulator*** and their mates, the ***demodulator*** and ***decoder***, which are the principal subjects of this book. These should be understood as a tandem, generating signals for large sets of messages, these signals being built from sequences of relatively simple modulator waveforms. An M-ary modulator is a device that accepts an input X_n selected from an alphabet of M characters and in response supplies one of M analog waveforms suitable for transmission over the physical channel. Essentially, the modulator is the interface to the actual communication medium and often involves such functions as frequency conversion, amplification, and antennalike transducers. (Communication engineers often separate the signal-building functions from the latter transmitter functions for reasons of flexibility of hardware and because the required engineering skills differ.)

The selection of the class of signals to be used, or the modulator design, is largely determined by the channel's anticipated noise, interference, and distortion characteristics. For example, if the channel imposes severe bandwidth limitations for the intended transmission rate, it is natural to use modulation methods that use special spectrum-shaping techniques to minimize distortion and to simplify decoding. In other cases, signal power is the precious commodity, and we will encounter design options that lower the required signal-to-noise ratio in exchange for greater bandwidth occupancy. Sometimes the design is influenced by the anticipated signal processing in the demodulator; for example, when the determination of the received signal's sinusoidal carrier phase reference angle is difficult (due to Doppler shifts, intentional jamming, or simply economic reasons), we should select modulation formats capable of being processed without the need for carrier phase references; that is, they can be "noncoherently" demodulated, as described in Chapter 3.

Three relevant examples of digital modulators will clarify their function, with mappings illustrated in Figure 1.1.2. These examples span a wide range of transmission rates and operating frequencies, but are common in their discrete-time, discrete-message-set nature.



(a) 300 bps frequency-shift-keying

(b) Alternate mark inversion as in *T1* transmission

(c) Optical 8-position PPM

Figure 1.1.2 Depiction of three modulation techniques.

Devices called modems[4] are used to connect various data terminals together via the public switched telephone network (PSTN). Older designs, now far from state of the art, transmitted at a rate of 300 bits/second, considered rapid for the time. A pulse of sinusoidal signal at audio frequency f_1 represents a 1 bit, while a sinusoid at audio frequency f_0 represents a 0, as depicted in Figure 1.1.2a. This application provides a special case of binary ($M = 2$) frequency shift keying. Today's telephone channel modems operate up to 90 times faster over the same medium, largely possible due to advanced modulation and coding techniques.

Another current system in North America (the T1 digital transmission system) sends binary information over coaxial cable or twisted-wire-pair media at 1.544 megabits per second using *alternate mark inversion*, wherein 0 is signaled by no current, and every

[4] Modem is a contraction for modulator/demodulator; also, a codec is a coder/decoder.

1 is represented by a current pulse, but with alternating polarity (Figure 1.1.2b). This modulator can supply $M = 3$ levels of current, with transmitter coding taking care of the alternating polarity. The reason for this curious alternation of polarities is to accommodate the inability of the channel to reproduce long strings of consecutive positive or negative polarity signals.

A third example pertinent to optical communication adopts a signaling interval, let's say 10 nanoseconds long, divided into eight time slots. Optical energy is radiated in exactly one slot per interval (by switching on a semiconductor laser source, say), and this modulation process is called pulse position modulation (PPM). This $M = 8$-ary modulator scheme seeks to communicate three bits of information per interval, or 300 megabits per second.

The ***channel encoder*** is a discrete-input, discrete-output device whose usual purpose is seen as providing some error-correction capability for the system. It does this by using a mapping from input sequences $\{U_n\}$ to code sequences $\{X_n\}$, which inserts ***redundancy*** and which utilizes ***memory***. Whereas in N modulator time slots, an uncoded system could transmit M^N possible signals, the coded system will enforce constraints that allow a smaller number of coded signals. In this sense, each modulator symbol doesn't carry as much information as it apparently could, and symbols are in some sense redundant. Memory is the other crucial aspect of good encoding schemes. In essence, a given message bit at the encoder input influences several, perhaps many, output symbols, hence waveform intervals. This provides a noise-averaging feature, which makes the decoder less vulnerable to the effects of noise, distortion, fading, and the like, occurring in one signaling interval. We will find that this is nothing more than an exploitation of the law of large numbers associated with a random channel mechanism.

An additional role of the channel encoder, although one less commonly attributed to it, may be that of spectral shaping. The memory of the encoder can, if desired, produce an output symbol stream that ultimately shapes the power spectrum of the signal produced by the modulator. An example is the alternate mark inversion technique described previously; a very simple channel encoder remembers the polarity of the previous 1 symbol and uses the opposite upon receiving the next 1. The resultant spectrum has very small power spectral content near zero frequency. Another important example is in coding for magnetic recording channels, where encoding the binary magnetization signal to satisfy run-length constraints helps to increase the information density per unit area of the medium.

We have seen that the combination of channel encoder and modulator provides a mapping from a bit stream to a signal waveform, which has aspects of redundancy and memory. In some applications, it is clear where these functions reside. For example, in the coding scheme used in the Voyager mission to the outer planets, message bits were stream encoded, producing two coded symbols for every message bit, with a memory length of six information bits. Each coded symbol then simply phase switched (by either 0° or 180°) the transmitted microwave signal. The memory and redundancy are clearly introduced by the encoder, and the modulator is a rather simple device.

In other schemes being used or studied today, the distinction is less clear. A principal example is that of continuous-phase modulation, to be studied in Chapter 6. Viewed most simply, this is just a phase-modulation process, which itself has memory, to enforce phase continuity, and perhaps even a higher degree of smoothness, upon the transmitted signal. This memory has a dramatic impact on the signal's power spectrum (important in satellite communication, mobile radio, and the like, where frequency congestion is high), but it is also clear that the process allows for noise averaging as well, since only certain phase patterns are allowed over a span of several intervals.

Other than making the organization of textbooks a bit more difficult, the issue of where to draw the demarcation line is not important. The classical view is that the encoder produces most all of the theoretically significant features (memory and redundancy), and the modulator, although essential, is generally uninteresting from a theoretical point of view. This view is still prevalent and in some cases proper, but the issue has become more clouded of late and properly so, for we should understand efficient signaling in a broader perspective.

Now that we have characterized the function of the channel encoder and modulator, we might expect that the demodulator and channel decoder ought to be easily understood. Often, however, a quite improper conclusion about optimal detection is made, that the demodulator should make its best judgment of what the modulator input symbol was in a given interval and then pass this decision on to the decoder, which in turn uses the known encoder structure to make a best judgment of what encoder input message was sent. This will be called the *error correcting code* viewpoint. Good encoder/decoder combinations can indeed overcome scattered, or even bursty, errors produced in the demodulator decision process and dramatically improve system performance. The fact is, however, much better performance is generally available if the demodulator resists the temptation to decide symbol by symbol, but instead passes to the decoder a sequence $\{Y_n\}$ containing likelihood information for the various modulator possibilities and lets the decoder merge all these pieces together to make message decisions. In the communication literature this latter process is often referred to as making *soft decisions*, that is, supplying only tentative estimates for the various symbols, instead of making an irrevocable (hard) decision on a modulator symbol. The term soft decision is somewhat misleading, however, for in efficient digital communication systems, the real *decisions* are the task of the channel decoder, not the demodulator.

This enlightened view is not a new one among theoreticians, but it is slow to penetrate the applications world. Certainly, in many situations it is simply inconvenient to provide anything but hard-decision symbols, particularly if the modem equipment already exists, and coding is an afterthought to improve performance. Furthermore, the soft-decision approach can be ineffective, if inappropriately handled, in certain impulsive noise situations, say due to pulse jamming. In any case, the essential notion is that the demodulator must operate in synergism with the decoder. Proper cooperation often provides enormous improvements in efficiency. Another way to state the case is that an efficient communication system is not obtained by first having the modem group build a good modulator/demodulator and then, when better performance is needed, turning to the coding experts for an add-on box (sometimes called a codec). The resulting system may indeed work better, although it is easy to produce cases where improperly applied coding techniques actually degrade performance. An integrated approach from the beginning is needed.

In the remainder of the book, we will examine in detail many different options for implementing the modulation and channel encoding functions just described and will analyze their performance in a variety of channel settings. First, however, we make a brief historical tour of the development of digital communications.

1.2 HISTORICAL NOTES

In the electrical context, digital communication has its roots in telegraphy. In 1838, Samuel F. B. Morse demonstrated a machine (the telegraph, meaning *distant writing*) for sending messages rapidly over long distances, which for the time was remarkable in itself. (Remember this was some 60 years prior to the demonstration of long-distance radio wave propagation.) More inteiesting perhaps is that this first high-speed electrical telecommunication device utilized simple current-on/current-off signaling to build a code, or discrete signal set, for sending alphanumeric characters. This code, later standardized as Morse code, used dots, or short current pulses, and dashes, or pulses typically about two times the duration of dots. The designers of this code put their good sense to work (long before information theory was born) and assigned short patterns to frequently used letters such as *e* (a single dot) and *t* (single dash). Numbers and punctuation marks, as well as rare letters such as *q*, *x*, and *z*, were assigned longer patterns. In this way, telegraphers could send typical messages in shorter time, and skilled operators could achieve throughputs of perhaps 75 words per minute, limited by the keying speed of the sender and the processing speed of the human decoder, but generally not by the transmission medium. Despite having its genesis in the digital framework, for about the next 100 years, communication became predominantly analog in the forms of the telephone (invented by Bell in 1872), radio, and television.

The work of Nyquist in the 1920s [5, 6], was perhaps the next major conceptual contribution to the development of digital communications. Nyquist developed the relationship between the available channel bandwidth and the maximum pulse transmission rate that would support zero interpulse interference and laid the foundation for the celebrated sampling theorem: a frequency-limited signal can be represented by the values of the signal taken at regular intervals if the sampling interval is sufficiently short, that is, $T_{\text{samp}} < (2B)^{-1}$, where B is the signal's bandwidth. This result allows communication to be at least a discrete-time process and is at the heart of most modern digital communication systems, where signals such as speech are sampled, quantized, digitally transmitted, and reconstituted into continuous-time signals for the user.

At about the same time, the seed for the modern view of information theory was planted by Hartley [7], who reasoned that information was related to the prior uncertainty about the message and that the amount of information contained was proportional to the logarithm of the number of possible messages. These ideas were not taken up again until the work of Shannon 20 years later.

The patent issued to Reeves in 1938 [8] marks another key step. Here, elementary (binary) sequences of fixed length, called pulse codes, were used to represent analog sample values by partitioning the range of analog values into small regions and transmitting the identification code for the region in question. This technique, known as pulse code modulation (PCM), occupies a central place today in transmission and storage of speech, music, visual and telemetry signals.

In the 1940s and 1950s, major advances were made along different theoretical lines, for example, in treating communication as a probabilistic process and using probability theory, which itself was rapidly maturing as a mathematical discipline, to optimize various aspects of the communication process. Fundamental was the development of stochastic noise theory [9] and procedures for optimal reception of messages [10–12]. Terms such as *expectation*, *maximum likelihood*, and *matched filter* became part of the communication theoretician's language.

The treatment of the communication process as a probabilistic process raised the issue of how efficient communication could ultimately be, given a probabilistic source and a probabilistic channel. In his landmark paper published in 1948 [1], Shannon showed there are fundamental limits to the throughput of information implied by the nature of the channel, but not on the reliability or accuracy, provided these throughput constraints are obeyed.[5] More specifically, Shannon showed there is a parameter attached to a channel, called *channel capacity*, denoted C, that has the significance that attempts at reliable communication are doomed to failure if the attempted digital information throughput exceeds C. More positively, though, we can achieve arbitrarily reliable communication if we signal at information rates less than C, where by reliable we mean "with an arbitrarily small message error rate." Shannon's demonstration of this result was quite remarkable, being an existence proof without showing the details of schemes that behave as described previously. The proofs indicate the need to send messages in long blocks of channel symbols to achieve some kind of noise averaging, but no constructive codes were proposed.

Shannon's discovery was truly a revolution in communication history (J. L. Massey [13] has likened it to the Copernican revolution), which overturned the former communication theory of how to send messages reliably, that is, send them several times, perhaps many times, and take some sort of majority vote. Of course, the throughput, or channel utilization, decreases with such increasing repetitions. Shannon showed it was not necessary to reduce the throughput to near zero in order to get reliable transmission, but only to less than the special number C already mentioned. Naturally, this better performance should not come without cost; we are faced with communicating using long blocks, and thus with presumably long delay (measured in signal intervals), and perhaps to using rather complicated schemes for encoding and decoding.

Around 1950, Golay [14] and Hamming [15] proposed the first nontrivial error-correcting codes, which utilized a number of redundant bits or parity check bits, along with the information bits, to achieve better performance than an uncoded system would achieve on the same channel. This, coupled with the existence of stronger codes shown to exist by Shannon, precipitated a wealth of subsequent work (see [16] and [17] for compendia of key early papers), and new modulation/coding techniques are still an important research topic.

In terms of system development, the T-carrier system initiated by the Bell System in the 1960s for digital telephony and video represents the first major penetration of digital transmission technology into a previously analog world. The Integrated Services Digital Network (ISDN) [18] represents a plan for a worldwide switched digital information

[5] It is perhaps clear that we have yet to formally define information throughput or rate; we will do so in Chapter 2.

network for voice, data, and visual information transfer. Satellite communication has been undergoing a steady transition from analog to digital transmission techniques. In late 1988, the TAT-8 transatlantic fiber-optic system was put into service, employing digital lightwave transmission. This system conveys about 40,000 simultaneous voice conversations. Presently, the vision is seamless communication of voice, data, and video over wide-area networks, using asynchronous transfer mode (ATM) digital packet architectures.

In more personal terms, we encounter digital communications technology everywhere today—at automatic teller machines, on Touch-Tone telephones, in listening to compact-disc recordings, in highly reliable computer memory systems, in communicating via cellular telephones, and in viewing the latest images from the outer planets. These developments have been enabled by many facets of emerging technology in addition to the communication-theoretic advances described in this text, and it must be said that many formerly academically interesting techniques have been made practical by huge advances in semiconductor technology.

1.3 OUTLINE OF BOOK

We begin our development in Chapter 2 with a survey of the basic probability and random process theory needed subsequently. Fundamental notions of decision theory are developed. Shannon's information theory and the concepts of entropy and channel capacity are introduced, along with simple demonstrations in the form of coding theorems that entropy and channel capacity are indeed important factors in establishing the possibility of reliable communication.

In Chapter 3, we begin the modulation and coding story by posing some simple, yet rather general channel models, which in various special cases represent many of the currently interesting channel scenarios. Following this, we formulate the signal-space view of communication theory, where signals and noise are represented in vector space terminology, as a means of deriving optimal receivers as well as providing a very useful visualization of the design problem. We will analyze all classical baseband and carrier modulation techniques, both with coherent and noncoherent detection, for the nonfading additive Gaussian noise channel as well as the Rayleigh fading channel. Our treatment here is in a classical vein; the emphasis is on the relationship between error probability, signal-to-noise ratio, and spectral bandwidth, assuming symbol-by-symbol, or uncoded, transmission. Spread spectrum modulation techniques are also discussed.

In Chapter 4, we move into the realm of coded transmission, beginning with a classification of important techniques. We then take up a more information-theoretic treatment of the communication process by considering, as did Shannon, the ensemble of possible codes. It is surprising that we can say very strong things about this set of codes, and thus demonstrate the existence of good codes, without ever having found one of them! The positive coding theorem is developed, demonstrating reliable communication at rates near the channel capacity C. In this discussion, the parameter R_0 emerges as a convenient single-parameter description of the coding potential of various systems, and we compute this parameter for different modulation choices, for different channels, and under differing detection strategies.

The emphasis shifts in Chapter 5 toward coding practice. Block coding is the topic, with the focus on cyclic binary and nonbinary codes. We describe the structural properties of Hamming, BCH, Reed–Solomon, and related codes, including maximum likelihood and algebraic decoding procedures for these codes. Code modification, code concatenation, and code interleaving are all treated as topics of practical importance. Performance analysis is given for the additive white Gaussian noise channel and the Rayleigh fading channel, emphasizing the impact of receiver quantization (hard versus soft decisions) and receiver side information for fading and interference channels.

Trellis coding is the topic of Chapter 6, beginning with convolutional codes. One principal attraction of trellis coding is the possibility of implementing optimal decoders with reasonable complexity, and we will study the decoding procedure known as the Viterbi algorithm in detail. The generating function approach for evaluating the performance of trellis codes is illustrated for the classical cases, such as coded phase-shift keying, and also for coded noncoherent communication. Newer trellis-coded schemes pioneered by Ungerboeck follow, along with the continuous-phase modulation techniques under this same unified framework. Threshold decoding and sequential decoding are studied briefly as suboptimal decoding procedures.

BIBLIOGRAPHY

1. Shannon, C. E., "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, July 1948, and vol. 27, pp. 623–656, October 1948.
2. Berger, T., *Rate Distortion Theory*. Englewood Cliffs, NJ: Prentice Hall, 1971.
3. Jayant, N. S., and Noll, P., *Digital Coding of Waveforms*. Englewood Cliffs, NJ: Prentice Hall, 1984.
4. Konheim, A., *Cryptography: A Primer*. New York: Wiley, 1981.
5. Nyquist, H. "Certain Factors Affecting Telegraph Speed," *Bell System Technical Journal*, vol., 3, no. 2, pp. 324–346, April 1924.
6. Nyquist, H., "Certain Topics in Telegraph Transmission Theory," *AIEE Transactions*, vol. 47, pp. 617–644, 1928.
7. Hartley, R. V. L., "Transmission of Information," *Bell System Technical Journal*, vol. 7, no. 3, pp. 535–563, July 1928.
8. Reeves, A. H., French Patent No. 851 183 on Pulse Code Modulation, October 3, 1938.
9. Rice, S. O., "Mathematical Analysis of Random Noise," *Bell System Technical Journal*, vol. 23, pp. 282–332, 1944.
10. North, D. O., "Analysis of the Factors That Determine Signal/Noise Discrimination," RCA Labs Report PTR-6C, June 1943, reprinted *Proceedings of IEEE*, vol. 51, July 1963.
11. Kotelnikov, V. A., *The Theory of Optimum Noise Immunity*. New York: McGraw-Hill, 1959. (Reprint of doctoral dissertation presented in Moscow, January 1947.)
12. Wiener, N., *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. Cambridge, MA: MIT Press, 1949.

13. Massey, J. L., Jr., "The Copernican Revolution of Communications," International Conference on Communications, Amsterdam, 1984, reprinted in *IEEE Communications Magazine*, December 1984, pp. 26–28.

14. Golay, M. J. E., "Notes on Digital Coding," *Proceedings of IEEE*, vol. 37, 1949.

15. Hamming, R. W., "Error Correcting and Error Detecting Codes," *Bell System Technical Journal*, vol. 29, pp. 147–160, 1950.

16. Berlekamp, E. R. (ed.), *Key Papers in the Development of Coding Theory*. New York: IEEE Press, 1974.

17. Slepian, D. (ed.), *Key Papers in the Development of Information Theory*. New York: IEEE Press, 1974.

18. *IEEE Communications Magazine*, March 1986, contains several articles devoted to the emergence of ISDN.

OTHER RECOMMENDED TEXTS ON DIGITAL TRANSMISSION

Anderson, J. B., Sundberg, C. E.-W., and Aulin, T., *Digital Phase Modulation*. New York: Plenum, 1986.

Benedetto, S., Biglieri, E., and Castellani, V., *Digital Transmission Theory*. Englewood Cliffs, NJ: Prentice Hall, 1987.

Bingham, J. A. C., *The Theory and Practice of Modem Design*. New York: Wiley-Interscience, 1988.

Blahut, R., *Digital Transmission of Information*. Reading, MA: Addison-Wesley, 1989.

Blahut, R., *Theory and Practice of Error Control Codes*. Reading, MA: Addison-Wesley, 1983.

Clark, G., and Cain, J. B., *Error Correction Coding for Digital Communications*. New York: Plenum, 1981.

Cover, T., and Thomas, J., *Elements of Information Theory*. New York: Wiley-Interscience, 1991.

Divsalar, D., Simon, M. K., Biglieri, E., and McLane, P., *Introduction to Trellis Coded Modulation and Applications*. New York: Macmillan, 1991.

Gallager, R., *Information Theory and Reliable Communication*. New York: Wiley, 1968.

Gitlin, R., Hayes, J., and Weinstein, S., *Data Communications Principles*. New York: Plenum Press, 1992.

Korn, I., *Digital Communications*. New York: Van Nostrand Reinhold, 1985.

Lee, E. A., and Messerschmitt, D. G., *Digital Communications*. Norwell, MA: Kluwer, 1988.

Lin, S., and Costello, D. J., *Error Control Coding: Fundamentals and Applications*. Englewood Cliffs, NJ: Prentice Hall, 1983.

McEliece, R. J., *The Theory of Information and Coding*. Reading, MA: Addison-Wesley, 1977.

Michelson, A., and Levesque, A., *Error Control Techniques for Digital Communication*. New York: Wiley, 1985.

Proakis, J., *Digital Communications*. New York: McGraw-Hill, 1989.

Sklar, B., *Digital Communications: Fundamentals and Applications*. Englewood Cliffs, NJ: Prentice Hall, 1987.

Sloane, N. J. A., and McWilliams, J., *The Theory of Error Correcting Codes*. Amsterdam: North Holland, 1977.

Viterbi, A. J., and Omura, J. K., *Principles of Digital Communications and Coding*. New York: McGraw-Hill, 1980.

Wozencraft, J., and Jacobs, I. M., *Principles of Communication Engineering*. New York: Wiley, 1967.

Ziemer, R., and Peterson, R., *Digital Communications and Spread Spectrum Systems*. New York: Macmillan, 1985.