# 1 Introduction

*Anytime*, *anywhere*, *anything* can be taken as the motto and objective of digital communications. *Anytime* means that one can communicate on a 24/7 basis; *anywhere* states this communication can take place in any geographical location, at minimum one no longer is tied to being close to one's land line; *anything* implies that not only traditional voice and video but also other messages can be transmitted over the same channel, principally text, and not only individually but in combination. In large part digital communication systems over the past three decades have achieved the three objectives. Text messaging in all its forms, such as email, internet access, etc., is a reality. Webcasting and podcasting are becoming common. All parts of the globe are connected to the world wide communication system provided by the Internet.

Perhaps, and arguably just as important, a fourth "any" can be added, *anybody*. Though perhaps not as well developed as the first three, digital communication has the potential to make communication affordable to everyone. One feature of digital circuitry is that its cost, relative to its capability, keeps dropping dramatically. Thus though analog communication could achieve the above objectives, *digital communications*, due to this increasingly low cost, flexibility, robustness, and ease of implementation, has become the preferred technology.

This text is an introduction to the basics of digital communication and is meant for those who wish a fundamental understanding of important aspects of digital communication design. Before detailing the text's content and organization, a brief history of digital communication and general comments about it are in order.

Though it does not appear on Maslow's hierarchy [1] of human needs, communication is crucial for any living organism, human or otherwise. To aid the process of communication, humans and human society have developed various techniques and technologies. The impetuses for the developments have been the need to increase the distances over which communication takes place, to increase the rate of communication, and to maintain the reliability of communication. It was only in the last half of the twentieth century that communication systems started to achieve reliable, universal, global communication. Many factors have contributed to this but at the core is the rapid adoption of digital communications.

*So what is digital communication?* Definitions vary, but the simplest one is that it is the communication (or transmission) of a message using a *finite alphabet* (symbol set) during a *finite time interval* (symbol interval). As such the raising of an eyebrow, a wink, the

nod of one's head or a shoulder shrug may be considered to be digital communications.[1] However, both the message set and the distances are limited in these examples. More meaningful digital communication systems were developed early: the Roman army used shields and the sun to flash signals over line-of-sight distances; similarly, North American natives used smoke signals, to list just two preindustrial digital communication systems. But communications as we know today and experience daily was ushered in during the nineteenth century. It started with the "harnessing" of electricity. This harnessing, which began in the mid-eighteenth century, meant that communication at distances further than one could see or hear became feasible.

In 1729, the English scientist Stephan Gray demonstrated that static electricity could be conducted by some materials, e.g., wet twine. The first idea for an electric telegraph was outlined in 1753 in a letter to the *Scots Magazine*. The letter was signed "C.M." and the anonymous author is believed to be Charles Morrison, a surgeon. He proposed the use of as many wires as there are letters in the alphabet. A message was sent sequentially in time with the specific letter indicated by sending an electrostatic charge over the appropriate wire. Dispatches could be sent at distances of 1 or 2 miles with considerable speed. Though there followed various variants and improvements on this idea, it was not until 80 years later that a practical digital communication system, the electric telegraph, was developed and patented. Indeed two patents were granted. One to Charles Wheatstone (of Wheatstone bridge fame) in 1837 in London; the other to Samuel Morse,[2] applied for in 1840, based on his 1837 caveat, with a US patent granted in 1849.

Until 1877, all rapid long distance communication depended upon the telegraph, in essence digital communication with only text messages being transmitted. But in 1877 the telephone was invented by Alexander Graham Bell and this heralded the arrival of long distance analog communications. Coupled with Hertz's discovery of the propagation of electromagnetic waves and Marconi's subsequent exploitation of this phenomenon to greatly increase communication distances, analog communication was ascendent for most of the twentieth century.

However, the second half of the twentieth century, particularly the last two decades, saw a resurgence in digital communications. In 1948 Claude Shannon published a landmark paper [2] in the annals of science in which he showed that by using digital communications it was possible even in the presence of noise to achieve a vanishingly small error probability at a finite communication rate (or finite bandwidth) and with finite signal power. His was a theoretical result and promised the Holy Grail that communication engineers have pursued since. At approximately the same time, R. W. Hamming proposed the Hamming codes [3] for error detection and correction of digital data. The invention of the transistor, also in 1948, and subsequent development of integrated circuitry provided the last component for a digital communications resurrection.

Initially digital communication systems were developed for deep space communications where data reliability was paramount and cost of lesser consideration. The first commercial

[1] In the context of a typical digital communication system block diagram presented later, it is left to the reader to identify the source alphabet, transmitter, receiver, channel, possible impairments, etc., in the described scenarios.

[2] Samuel Morse also devised the dot-dash system known as Morse code.

application started in 1962 when Bell Systems introduced the T1 transmission system for telephone networks.

However, analog communication was still dominant and the first mobile telephone system introduced in North America in the 1980s was analog based. But the ever increasing integrated chip densities and the concomitant decrease in cost meant that the intensive signal processing required by digital communications became feasible. And indeed the late 1980s and the last decade of the twentieth century saw several digital communication systems developed. The GSM mobile phone system[3] was introduced in Europe and DARPA's (Defense Advanced Research Projects Agency) sponsor of a computer communication network, led to the establishment of the Internet. At the end of the millennium one could reasonably state that digital communication systems were dominant again.

*So what is a digital communication system and what does it involve?* Let us start to answer these questions by considering two general paradigms.

## 1.1 Open system interconnection (OSI) model

The OSI model was developed explicitly for computer communications over public communication networks. However, it may also serve as a model for other communications. Communication networks such as the public switched telephone network (PSTN) and the Internet are very complex systems which provide transparent and seamless communication between cities, different countries, different languages, and cultures. The OSI model serves to abstract the fundamentals of such a system. It is a seven-layer model for the functions that occur in a communication process.

Figure 1.1 illustrates the different layers. Each layer performs one or a number of related functions in the communication process. Using terminology that arose from computer communications, the seven layers, or any subset of them, are often called a protocol stack. Layers at the same level are known as peer processes. The following description of the functionality of each layer is taken from [4].

(1) *Physical layer* This first layer provides a physical mechanism for transmitting bits between any pair of nodes. The module for performing this function is often called a *modem* (*mo*dulator and *dem*odulator).
(2) *Data link layer* This second layer performs error correction or detection in order to provide a reliable error-free link to the higher layers. Often, the data link layer will retransmit packets that are received in error, but in some implementations it discards them and relies on higher layers to do the retransmission. The data link layer is also responsible for the ordering of packets, to make sure that all packets are presented to the higher layers in the proper order. The role of the data link layer is more complicated when multiple nodes share the same media, as usually occurs in wireless systems. The

[3] The abbreviation "GSM" originally came from the French phrase *Groupe Spécial Mobile*. It is also interpreted as the abbreviation for "Global System for Mobile communications."

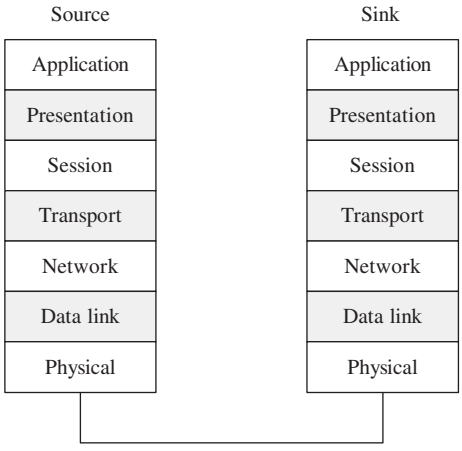


**Fig. 1.1** Illustration of different layers in the OSI model.

component of the data link layer that controls *multiple-access communications* is the *medium access control* (MAC) *sublayer*, the purpose of which is to allow frames to be sent over the shared media without undue interference from other nodes.

(3) *Network layer* This third layer has many functions. One function is to determine the *routing* of the packet. A second is to determine the *quality of service* (QoS), which is often controlled by the choice of a connectionless or connection-oriented service. A third function is *flow control*, which ensures that the network does not become congested. Note that the network layer can generate its own packets for control purposes. Often, there is a need to connect different subnetworks together. The connectivity is accomplished by adding an *internet sublayer* to the network layer – a sublayer that provides the necessary translation facilities between the two networks.

(4) *Transport layer* The fourth layer separates messages into packets for transmission and reassembles the packets at the other end. If the network layer is unreliable, the transport layer provides reliable end-to-end communications by retransmitting incomplete or erroneous messages; it also restarts transmissions after a connection failure. In addition, the transport layer may provide a multiplexing function by combining sessions with the same source and destination; an example is parallel sessions consisting of a gaming application and a messaging application. The other peer transport layer would separate the two sessions and deliver them to the respective peer applications.

(5) *Session layer* This fifth layer finds the right delivery service and determines access rights.

(6) *Presentation layer* The major functions of the presentation layer are data encryption, data compression, and code conversion.

(7) *Application layer* This final layer provides the interface to the user.

Though the OSI model has conceptual utility, one should realize that in practice, for many (if not most) digital communication systems, use of the OSI model becomes quite ambiguous if not downright misleading. To give a simple but important example consider

error correction or detection. Though assigned to the data link layer, it is just as readily found at the physical layer when hardware error correction/detection integrated circuits are designed. Even more dramatically, data encryption/decryption, data compression, and code conversion can be performed at the physical layer level. Therefore, though the OSI model is of use to elucidate the different functions, in practical implementations of a digital communication system the levels are blurred considerably. For engineering purposes a more pertinent block diagram of a digital communication system is discussed in the next section. In essence, such a block diagram is mainly concerned with the functionality and issues in the *physical layer* of the OSI seven-layer model.

## 1.2 Block diagram of a typical digital communication system

Figure 1.2(a) shows a block diagram applicable to either an analog or a digital communication system (to be precise, the "synchronization" block is typically only needed in a digital system). It is rare for the system designer to have control over the channel and even rarer, if ever, for the designer to have control over the source. Therefore design and analysis are focused on the transmitter and receiver blocks. The design, of course, must take the source and channel characteristics into consideration. With regard to the transmitter and receiver blocks, for a digital communication system they can be further subdivided for our purposes as shown in Figure 1.2(b).

Consider the transmitter block. The source is typically first passed through a source encoder which prepares the source messages. The encoder details depend on the source and may be further subdivided. For a voice or video source it would consist of, at minimum, an analog-to-digital converter, for a keyboard it could be an ASCII code mapper, etc. Data encryption may be considered to be part of it. Ideally what the source encoder should do is remove all redundancy from the source message, represent it by a symbol drawn from a *finite alphabet*, and transmit this symbol every $T_s$ seconds. Typically the alphabet is binary and the source encoder output is a bit stream or sequence. Removal of the redundancy implies that the source rate (typically measured in bits/second) is reduced, which in turn, as shall be seen, means that the required frequency spectrum bandwidth is reduced.

Having removed at least some of the redundancy by means of the source encoder, it may appear to be counterintuitive to have redundancy added back in by the channel encoder. This redundancy, however, is added back in a *controlled* fashion for error detection/correction purposes.[4] The channel encoder therefore maps the input symbol sequence into an output symbol sequence. To transmit this symbol sequence across a physical channel requires energy and this is the function of the modulator block. It takes the symbol occurring in each $T_s$ (symbol interval) and maps it onto a *continuous-time* waveform which is then sent across the channel. It is important to emphasize that, over any finite

[4] The reader may be familiar with odd/even parity check bits which add an extra binary digit for error detection purposes at the receiver.

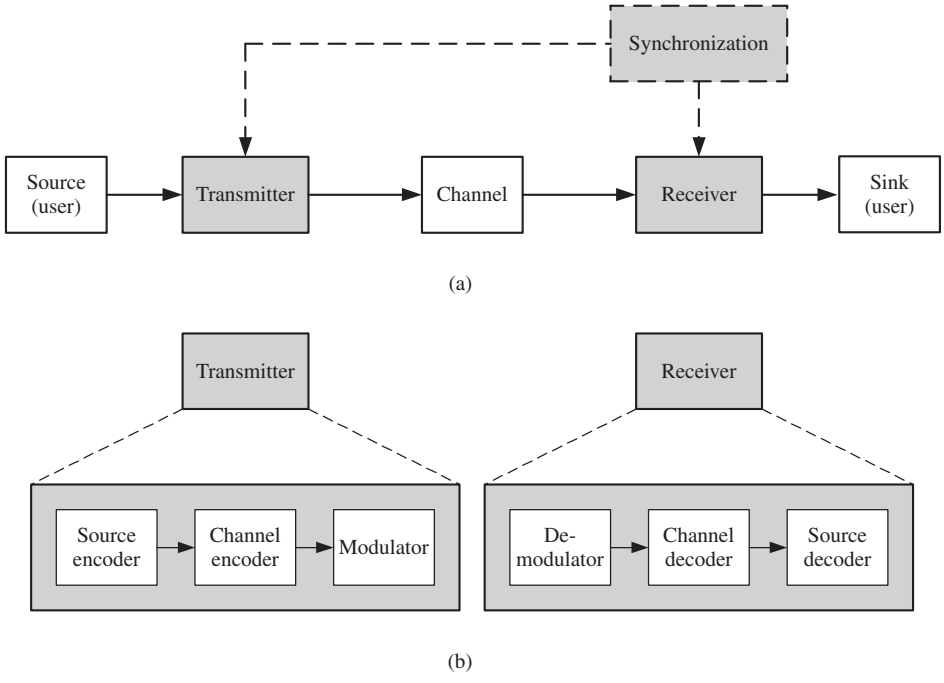


**Fig. 1.2** (a) General block diagram of a communication system. The "synchronization" block is only present in a digital system. (b) More detailed block diagram of the transmitter and receiver blocks of a digital communication system. Note that the outputs of the source and channel encoders are *digital* data streams, while that of the modulator block is a *continuous-time* signal.

time interval, the continuous-time waveform at the output of a digital communication system belongs to a *finite set* of possible waveforms. This is in sharp contrast to the analog communication system, in which the modulator directly maps the analog message to a continuous-time signal for transmission. Since analog messages are characterized by data whose values vary over a continuous range, the output of an analog communication system can assume an *infinite* number of possible waveforms. As shall be seen throughout the text, it is the finite property of the sets of the digital messages and modulated waveforms that makes the digital communication system more reliable than its counterpart by applying advanced signal processing algorithms at both the transmitter and the receiver.

The subdivision of the transmitter block, though still relevant, is somewhat classical. The channel encoder/modulator blocks can be, and are, combined to produce a coded modulation paradigm. Indeed all three blocks, source encoder, channel encoder, and modulator, can be combined though this approach is still in the research stage.

At the receiver, one simply passes the received signal through the inverse of the operations at the transmitter; simply, except it is not that simple due to the influence of the channel. If the channel did not filter or distort the signal, did not add noise to it, and if there was no interference from other users, then it would be simple. However, some or all

of these degradations are present in physical channels. Therefore one must attempt to overcome the degradations with one's design. For the design or analysis one needs reasonable engineering models of the channel. The channel model depends on the media over which the transmission of modulated signals takes place. Consider guided media such as twisted-pair wire, coaxial cable, and optical fiber. In these the background noise is mainly Gaussian. However, they may exhibit a signal distortion where the transmitted signal is "smeared" out and causes intersymbol interference, i.e., signals in adjacent or nearby time slots interfere with each other. In twisted-pair wire and coaxial cable the smearing occurs due to the finite bandwidth property, while in optical fiber it is due to the dispersion of the light as it travels down the fiber.[5]

A space channel, i.e., satellite communications, typically only adds Gaussian noise to the received signal. The terrestrial microwave channel is similar but the transmitted signal may also be subjected to reflection, diffraction, and refraction, which leads to fading. Fading is a predominant degradation in mobile communications where the signal path from the transmitter to the receiver changes rapidly. In mobile communications, because a great many users must be accommodated in a given geographic area, interference from other users becomes a factor and the chosen modulation/demodulation must reflect this.

As mentioned above, the designer of a communication system typically has little or no control over the source or channel. One is primarily concerned with the transmitter/receiver design. In this introductory text, the major concern is with the *mo*dulator/*dem*odulator blocks, commonly called the *modem*, which are part of every digital communication system. The introduction is concluded by an outline of the text contents.

## 1.3 Text outline

Since digital communications, particularly at the physical layer and particularly the modem block, involves the transmission of continuous-time waveforms with their reception corrupted by continuous-time noise signals, the next two chapters are concerned with continuous-time signals and systems. Chapter 2 deals with deterministic signals and different methods of characterizing and analyzing them. The material covered is that found in a typical undergraduate signals/systems course. As such it is meant primarily for review purposes but since it is quite self-contained it may be used by a reader with little or no background in the subject. The reader is encouraged to read it and also to attempt the problems at the chapter's end, or at least to read them.

Similar comments apply to Chapter 3 which deals with random signals, or random processes as common terminology terms them (stochastic processes is another name). First, a review of probability theory and random variables is provided. A random process is in essence nothing more than a random variable that evolves in time. Though one initially encounters a random process as additive noise in the channel model it should be pointed

[5] The bandwidth of twisted-pair wire is on the order of kilohertz, that of coaxial cable on the order of megahertz, and for fiber optic it is gigahertz.

out that basically any message source is best modeled as a random signal. Furthermore, the time-varying characteristic of a wireless channel can also be modeled by a random process.

Chapter 4 starts the reader on the road of digital communication. It deals with what, arguably, are the two most important sources, audio and video. These are sources whose outputs are inherently analog in nature. The chapter investigates fundamental principles of converting the analog signal into a digital format. The techniques are applicable to any analog source, e.g., a temperature measurement system. Converting the analog signal to a digital format can be considered to be an elementary form of source encoding.

Chapter 5 is where the study of digital communication systems starts in earnest. Since digital communication systems are found in diverse configurations and applications, the chapter develops a general approach to analyze and eventually design modems. The underlying philosophy is based on the phrase "*short-term pain for long-term gain*".[6] First the modulator output and channel noise are characterized. Then the demodulator is designed to satisfy a chosen performance criterion, namely to minimize the probability of (symbol) error. This is in contrast to analog communications where the criterion is typically the power signal-to-noise ratio (SNR) at the receiver output. The two criteria are related but are not synonymous. Though the symbols are assumed to be *bi*nary digi*t*s (bits) the concepts are readily extended to nonbinary symbols, and they are in later chapters. The chapter concludes with the derivation of the power spectral density (watts/hertz) for general memoryless binary modulation.

Chapters 6 and 7 deal with modem design. Chapter 6 considers baseband signaling. Loosely speaking, in baseband modulation the transmitted signal energy is concentrated around 0 hertz in the frequency band. However, the occupied band can be quite large, ranging from kilohertz (telephone line) to megahertz (cable) and gigahertz (fiber). Since using electromagnetic radiation is not feasible, baseband communication is over guided media. Chapter 7 considers modulation where the modulated signal is shifted to an appropriate frequency band (passband modulation) and then typically transmitted via antennas. Typically the channels are those found in space communications, terrestrial microwave communications, and mobile communications. However, passband modulation may also be found in cable channels for instance.

The analysis/design is concerned with the bandwidth and the average power (or equivalently energy) required by a given or proposed modulation format to achieve a certain performance level, as measured by the bit error probability. Bandwidth and power are viewed as the two natural resources that one has to utilize in the design. To reduce the bandwidth requirement at a given average power-per-symbol level (and hence error performance) one can resort to a nonbinary or $M$-ary modulation format. This is the topic of Chapter 8.

[6] Though the phrase sounds good, caution is advised. Excluding its application to dental visits, the second author is reminded of a political scenario where a federal minister proposed a budget that included a 25 cents tax increase on a liter of petrol and invoked the phrase. Needless to say the budget and government were defeated on a motion of confidence. In these days of global warming perhaps the minister was correct, though that was hardly his motivation. Furthermore, as subsequent events unfolded what resulted was the nation went through long-term pain for no gain (but this is only a personal opinion).

Up to Chapter 8 the channel model is that of additive white Gaussian noise (thermal noise) with a bandwidth that is infinite or at least sufficiently large so that any effect it has on the transmitted signal can be ignored. Chapter 9 studies the effect of bandlimitation. Bandlimitation results in interference where the transmitted signal in a given signaling interval interferes with the signals in adjacent intervals. Therefore, besides random thermal noise, one has to contend with intersymbol interference. Methods to eliminate or mitigate this interference are investigated in this chapter.

Wireless communications involves a transmitted signal being received over many paths. This results in a phenomenon called fading where the received signal strength varies with time. Chapter 10 develops this important channel model and the various challenges and solutions associated with it.

The modulation/demodulation aspect concludes with Chapter 11 where three modern modulation approaches are described and analyzed. The first modulation, known as trellis-coded modulation (TCM) was discovered in the late 1970s and developed during the 1980s. Its main feature is that a significant bandwidth reduction is achieved with no power or error penalty. The second modulation paradigm presented is code-division multiple access (CDMA) used in the so-called third (and future) generations of wireless communication networks. Finally space-time coding, introduced in the late 1990s, concludes the chapter. Space-time coding provides a significant improvement for wireless communication where fading is present. Though advanced, the modulations are quite readily understood in terms of the fundamental background material presented in the text up to this chapter. In fact all three modulations can be classed under the rubric of coded modulation where coding/modulation is viewed as a single entity. The coding (i.e., adding redundancy), however, is performed differently for the three modulations: in TCM it is done in the time domain, in CDMA in the frequency domain, and in space-time modulation the redundancy is accomplished by having a symbol and its variants transmitted over multiple antennas.

The book concludes with a chapter on synchronization, which appears last not because it is any less important than the other topics covered but simply because one has to start and end somewhere. Synchronization provides all the timing necessary in a digital communication system and therefore is of equal importance to modem design. Chapter 12 presents two very common circuits used in digital communications: the phase-locked loop and the early–late gate.

## References

[1] A. H. Maslow, "A theory of human motivation," *Psychological Review*, vol. 50, pp. 370–396, 1943.

[2] C. E. Shannon, "A mathematical theory of communications," *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–657, 1948.

[3] R. W. Hamming, "Error detecting and error correcting codes," *Bell System Technical Journal*, vol. 29, pp. 147–160, Apr. 1950.

[4] S. Haykin and M. Moher, *Modern Wireless Communications*. Prentice Hall, 2005.

# 2 Deterministic signal characterization and analysis

## 2.1 Introduction

The main objective of a communication system is the reliable transfer of information over a channel. Typically the information is represented by audio or video signals, though one may easily postulate other signals, e.g., chemical, temperature, and of course text, i.e., the written word which you are now reading. Regardless of how the message signals originate they are, by their nature, best modeled as a random signal (or process). This is due to the fact that any signal that conveys information must have some uncertainty in it. Otherwise its transmission would be of no interest to the receiver, indeed the message would be quite boring (known knowns so to speak[1]). Further, when a message signal is transmitted through a channel it is inevitably distorted or corrupted due to channel imperfections. Again the corrupting influences such as the addition of the ever present thermal noise in electronic components, the multipath fading experienced in wireless communications, are unpredictable in nature and again best modeled as nondeterministic signals or random processes.

However, in communication systems one also utilizes signals that are deterministic, i.e., completely determined and therefore predictable or nonrandom. The simplest example is perhaps the carrier used by AM or FM analog modulation. Another common example is the use of test signals to probe a channel's characteristics. Channel imperfections can also be modeled as deterministic phenomena: these include linear and nonlinear distortion, intersymbol interference in bandlimited channels, etc.

This chapter looks at the characterization and analysis of deterministic signals. Random signals are treated in the next chapter. It should be mentioned at the outset that in certain situations whether a signal is considered deterministic or random is in the eyes of the viewer. Only experience and application can answer this question. The chapter reviews deterministic signal concepts such as Fourier series, Fourier transform, energy/power spectra, auto and crosscorrelation operations. Most of the concepts discussed carry over to random signals. Though intended to be self-contained it is expected that the reader has been exposed to these concepts in other undergraduate courses, hence the treatment is presented succinctly.

[1] For a discussion of knowns and unknowns see "The Secret Poetry of Donald Rumsfield" website http://maisonbisson.com/blog/post/10086/.