Cambridge University Press
978-0-521-78014-8 - Modem Theory: An Introduction to Telecommunications
Richard E. Blahut
Excerpt
More information

1 Introduction

A point-to-point communication system transfers a message from one point to another through a noisy environment called a *communication channel*. A familiar example of a communication channel is formed by the propagation of an electromagnetic wave from a transmitting antenna to a receiving antenna. The message is carried by the time-varying parameters of the electromagnetic wave. Another example of a communication channel is a waveform propagating through a coaxial cable that connects a jack mounted on an office wall to another such jack on another wall or to a central node. In these examples, the waveform as it appears at the receiver is contaminated by noise, by interference, and by other impairments. The transmitted message must be protected against such impairments and distortion in the channel. Early communication systems were designed to protect their messages from the environment by the simple expedient of transmitting at low data rates with high power. Later, message design techniques were introduced that led to the development of far more sophisticated communication systems with much better performance. Modern message design is the art of piecing together a number of waveform ideas in order to transmit as many bits per second as is practical within the available power and bandwidth. It is by the performance at low transmitted energy per bit that one judges the quality of a digital communication system. The purpose of this book is to develop modern waveform techniques for the digital transmission of information.

1.1 Transmission of information

An overview of a digital communication system is shown in Figure 1.1. A message originating in an information source is to be transmitted to an information user through a channel. The digital communication system consists of a device called a *transmitter*, which prepares the source message for the communication channel, and another device called a *receiver*, which prepares the channel output for the user. The operation of the transmitter is called *modulation* or *encoding*. The operation of the receiver is called *demodulation* or *decoding*. Many point-to-point communication systems are two-way

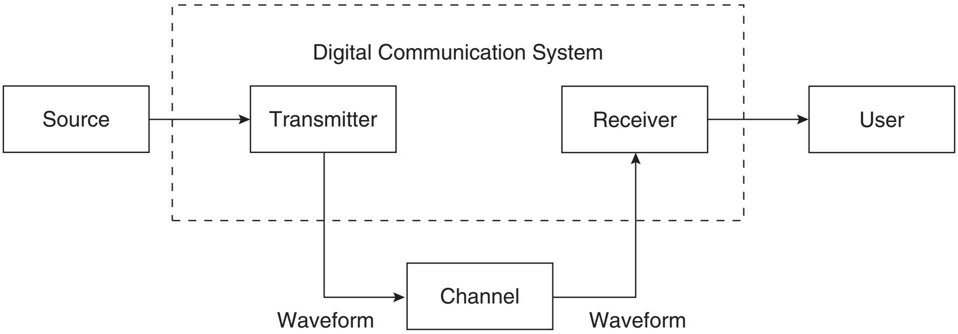

Figure 1.1. Overview of a digital communication system.

systems in which both a modulator and a demodulator are combined in a single package called a *modem*, the theory of which is the topic of this book.

At the physical level, a communication channel is normally an analog channel in that it transmits waveforms. The user information may arise as digital source data or may arise as an analog signal. The source data must eventually be modulated into an analog waveform that suits the analog channel. When the source signal is an analog waveform – perhaps a continuous-time analog waveform such as voice or video – a digital communication system first converts that waveform to a digital representation (which may take the form of a stream of bits), processes that digital representation in some way, but then converts it back into a continuous-time analog waveform for passage through the channel. The analog waveform passing through the channel will be completely different from the analog waveform generated by the source. The analog waveform at the channel output may be converted later by means of sampling and quantization to intermediate digital data for processing and demodulation. The intermediate digital data may be very different in its character from both the digital data that was transmitted and the final digital data produced by the demodulator. Ultimately, if required by the user, the digital data may be reconverted to its original analog form, such as voice or video. This multiple conversion between digital data and analog data may seem to be complicated and expensive, but it is worthwhile for many reasons. The channel waveform is matched to the nature of the channel, not to the nature of the source. Moreover, the digital data can be rerouted through many kinds of digital links or storage devices while it is in digital form, or it can be merged and mingled with other data traffic passing through a network.

Analog modulation is still widely used in radio and television, and until recently, was also used in phonography and voice telephony. Analog modulation techniques make relatively superficial changes to the signal in order to send it through the channel; there is no significant effort to tailor the waveform to suit the channel at any deeper level. Digital communication waveforms are more sophisticated. Digital communication theory endeavors to find waveforms that are closely matched to the characteristics of the channel and that are tolerant of the impairments in the channel so that the reliable flow of information through the channel is ensured. The characteristics of the source are of no interest in designing good waveforms for a channel. Good waveforms for digital communication are designed to match the characteristics of the channel; the source information is then encoded into this channel waveform. A digital communication system might require a considerable amount of electronic circuitry to translate the source waveform into a form more suitable for the channel, but electronics is now cheap. In contrast, most channels are comparatively expensive, and it is important to make the best use of a channel.

Even an application that might appear to be intrinsically an analog application, such as broadcast television, can be partitioned into two tasks – the task of delivering so many bits per second through the channel and the task of representing the video signal by the available number of bits per second. This is the important and (once) surprising separation principle of information theory, which says that the task of transmitting the output of a source through a channel can be separated, without meaningful loss, into the task of forming a binary representation of the source output and the task of sending a binary datastream through the channel. For digital transmission to be effective, both of these tasks must be implemented efficiently. Otherwise, there will be disadvantages such as increased bandwidth or larger transmitted power. The source data must first be compressed, then modulated into a suitable transmission waveform, perhaps a spectrally-efficient waveform that carries multiple bits per second per hertz, or perhaps an energy-efficient waveform that uses low power but occupies a large bandwidth.

The disadvantages that were once cited for digital communications are really not compelling; any validity that these claims once had has crumbled under the progress of technology. Cost was once a significant disadvantage of digital communication systems, but is no longer. Digital modulation of an analog signal was once believed to require larger bandwidth than direct analog modulation. This is not true. By using modern methods of data compression, data compaction, and bandwidth-efficient modulation – conveying multiple bits per second per hertz – a digital communication system will actually use less bandwidth than an analog communication system. Another presumed disadvantage that is sometimes mentioned is that of quantization noise. Quantization noise, however, is completely under the control of the designer of the quantization scheme and will be the only important source of noise in the signal that is presented to the user. The modern view is that quantization noise is a price cheerfully paid for the more important advantage of removing the effects of channel noise and other impairments from the received signal. It is a truism of information theory that, in a well-designed system, the quantization noise will always be less than the channel noise it replaces.

On the other hand, there are numerous and compelling advantages of digital communication. Every link becomes simply a “bit pipe” characterized by its data rate and its probability of bit error. It makes no difference to the communication links whether the transmitted bits represent a digitized voice signal or a computer program. Many kinds of data source can share a common digital communication link, and the many kinds of communication links are suitable for any data source. Errors due to noise and interference are almost completely suppressed by the use of specialized codes for the prevention of error. A digital datastream can be routed through many physically different links in a complex system, and can be intermingled with other digital traffic in the network. A digital datastream is compatible with standardized encryption and antijam equipment.

The digital datastream can be regenerated and remodulated at every repeater that it passes through so that the effect of additive channel noise, or other impairments, does not accumulate in the signal. Analog repeaters, on the other hand, consist of amplifiers that amplify both signal and noise. Noise accumulates in an analog communication waveform as it passes through each of a series of repeaters.

Finally, because digital communication systems are built in large part from digital circuitry, data can be readily buffered in random-access digital memories or on magnetic disks. Many functions of a modem can be programmed into a microprocessor or designed into a special-purpose digital integrated circuit. Thus a digital data format is compatible with the many other digital systems and subsystems of the modern world.

1.2 A brief historical survey

The historical development of modem theory can be divided into several phases: traditional methods such as PSK, PAM, QAM, and orthogonal signaling, which were developed early; the multilevel signal constellation designs of the 1970s; the coded modulation and precoding techniques of the 1980s and 1990s, and the graphical methods of the past decade. It is a curiosity of technological history that the earliest communication systems such as telegraphy (1832,1844) actually can be classified as digital communication systems. Even the time-honored Morse code is a digital communication waveform in that it uses a discrete alphabet. Telegraphy created an early communications industry but lacked the popular appeal of later analog communication systems such as the telephone (1876) and the phonograph (1877). These analog communication systems were dominant for most of the twentieth century.

The earliest broadcast systems for communication were concerned with the transfer of analog continuous signals, first radio signals (1920) and then television signals (1923, 1927). Analog modulation techniques were developed for embedding a continuous-time signal into a carrier waveform that could be propagated through a channel such as an electromagnetic-wave channel. These techniques are still employed in systems that demand low cost or have strong historical roots, such as radio, telephony, and television. There are indications, however, that analog modulation is becoming outdated even for those applications, and only the enormous investment in existing equipment will forestall the inevitable demise of the familiar AM and FM radios. Indeed, the evolution from digital communication to analog communication that began in the 1870s is now being countered by an evolution from analog communication back to digital communication that found its first strength in the 1970s. Even the analog phonograph record, after 100 years of popularity and dominance, has now been completely superseded by the compact disk.

The earliest radio transmitters used a form of analog modulation called *amplitude modulation*. This method maps a signal $s(t)$ into a waveform $c(t)$ given by

$$c(t) = [1 + ms(t)] \cos 2\pi f_0 t$$

where m is a small constant called the *modulation index* such that $ms(t)$ is much smaller than one, and f_0 is a constant called the *carrier frequency* such that f_0 is large compared to the largest frequency for which $S(f)$, the Fourier transform of $s(t)$, is nonzero. Even though the fidelity of the received signal is not noteworthy, amplitude modulation became very popular early on because the mapping from $s(t)$ to $c(t)$ could be implemented in the transmitter very simply, and the inverse mapping from $c(t)$ to $s(t)$ could be implemented simply in the receiver, though only approximately.

Frequency modulation is an alternative analog modulation technique given by the following map from $s(t)$ to $c(t)$:

$$c(t) = \sin\left(2\pi f_0 t + \int_0^t ms(\xi)d\xi\right)$$

where, again, the carrier frequency f_0 is large compared to the largest frequency for which $S(f)$ is significant. Frequency modulation was naively proposed very early as a method to conserve the radio spectrum. The naive argument was that the term $ms(t)$ is an "instantaneous frequency" perturbing the carrier frequency f_0 and, if the modulation index m is made very small, the bandwidth of the transform $C(f)$ could be made much smaller than the bandwidth of $S(f)$. Carson (1922) argued that this is an ill-considered plan, as is easily seen by looking at the approximation

$$c(t) \approx \sin 2\pi f_0 t + \cos 2\pi f_0 t \left[m \int_0^t s(\xi)d\xi\right]$$

when m is small. The second term has the same Fourier transform as the bracketed component, but translated in frequency by f_0. Because the integral of $s(t)$ has the same frequency components as $s(t)$, the spectral width is not reduced. As a result of this observation, frequency modulation temporarily fell out of favor. Armstrong (1936) reawakened interest in frequency modulation when he realized it had a much different property that was desirable. When the modulation index is large, the inverse mapping from the modulated waveform $c(t)$ back to the signal $s(t)$ is much less sensitive to additive noise in the received signal than is the case for amplitude modulation, – at least when the noise is small. Frequency demodulation implemented with a hardlimiter suppresses noise and weak interference, and so frequency modulation has come to be preferred to amplitude modulation because of its higher fidelity.

The basic methods of analog modulation are also used in modified forms such as single-sideband modulation or vestigial-sideband modulation. These modified forms are attempts to improve the efficiency of the modulation waveform in its use of the spectrum. Other analog methods, such as Dolby (1967) modulation, are used to match the analog source signal more closely to the noise characteristics of the channel. All methods for modifying the techniques of analog modulation are stopgap methods. They do not attack the deficiencies of analog modulation head-on. Eventually, possibly with a few exceptions, such methods will be abandoned in favor of digital modulation.

The superiority of digital communication seems obvious today to any observer of recent technological trends. Yet to an earlier generation it was not obvious at all. Shannon's original development of information theory, which was published in 1948 and implicitly argued for the optimality of digital communications, was widely questioned at the time. Communication theory became much more mathematical when Shannon's view became widely appreciated. His view is that communication is intrinsically a statistical process, both because the message is random and because the noise is random. The message is random because there is little point in transmitting a predetermined message if the receiver already knows it. If there are only a few possible predetermined messages already known to the receiver, one of which must be sent, then there is no need to send the entire message. Only a few prearranged bits need to be transmitted to identify the chosen message to the receiver. But this already implies that there is some uncertainty about which message will be the chosen message. Even this simple example introduces randomness, and as the number of possible messages increases, the randomness increases, and so the number of bits needed in the message increases as well.

Randomness is an essential ingredient in the theory of communication also because of noise in the channel. This statistical view of communication, encompassing both random messages and noisy channels, was promoted by Shannon (1948, 1949). Earlier, Rice (1945) had made extensive study of the effect of channel noise on received analog communication waveforms. Shannon developed the broader and (at that time) counterintuitive view that the waveform could be designed to make the channel noise essentially inconsequential to the quality of the received waveform. He realized that combating noise was a job for both the transmitter and the receiver, not for the receiver alone. In his papers, Shannon laid a firm foundation for the development of digital communication. A different paper dealing with applications that were transitional between analog and digital communication was due to Oliver, Pierce, and Shannon (1948). The period of the 1940s appears to be the time when people began thinking deeply about the fundamental nature of the communication problem and the return to digital signaling began to accelerate. There were, however, many earlier studies and applications of digital signaling as in the work of Nyquist (1924, 1928) and Hartley (1928). Aschoff (1983) gives a good early history of digital signaling.

1.3 Point-to-point digital communication

A simple block diagram of a point-to-point digital communication system is shown in Figure 1.1. The model in Figure 1.1 is quite general and can be applied to a variety of communication systems, and also to magnetic and optical storage systems. The boxes labeled "channel," "source," and "user" in Figure 1.1 represent those parts of the system that are not under the control of the designer. The identification of the channel may be somewhat arbitrary because some of the physical components such as amplifiers might in some circumstances be considered to be part of the channel and in other circumstances might be considered to be part of the modulator and demodulator.

It is the task of the designer to connect the data source to the data sink by designing the boxes labeled "transmitter" and "receiver". These boxes are also called, more simply, the *encoder* and *decoder* or the *modulator* and *demodulator*. The latter names are usually preferred when the channel is a waveform channel while the former are usually preferred for discrete channels. Consequently, the transmitter is also called the *encoder/modulator*, and the receiver is also called the *demodulator/decoder*. Often a single package that can be used either as a transmitter or a receiver (or both simultaneously) is desired. A modulator and demodulator combined into a single box is called a *modem*. The term "modem" might also be used to include the encoder and decoder as well, and usually includes other supporting functions that extend beyond modulation and demodulation but which are needed to make the modem work. The terms "transmitter" and "receiver" are sometimes preferred as the broader terms that include such supporting functions. Figure 1.2 and Figure 1.3 show the functions normally included in the transmitter and receiver.

Modern practice in the design of communication systems is to separate the design tasks associated with the data source and the data user from the design tasks associated with the channel. This leads technology in the direction of greater flexibility in that the source data, when reduced to a stream of bits, might be transmitted through any one of many possible channels or even through a network of different channels. To

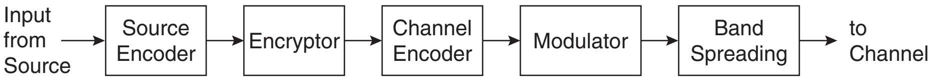

Figure 1.2. Primary transmitter functions.

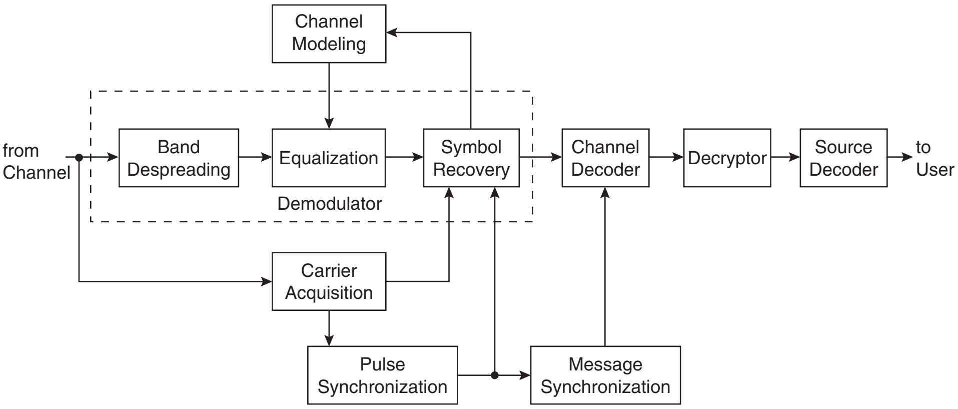

Figure 1.3. Primary receiver functions.

do this, the functions of the transmitter and receiver are broken into more detailed functions as described by the block diagrams of Figures 1.2 and 1.3. The transmitter includes a source encoder, a channel encoder, and a modulator; the receiver includes a demodulator, a channel decoder, and a source decoder. Information theory teaches us that there is no consequential loss in performance because of partitioning the problem in this way. Moreover, for our topics, there is no loss in generality if the interface between the source encoder and the channel encoder, as well as the interface between the channel decoder and source decoder are regarded to be serial streams of binary data.

The source data entering a digital communication system may be analog or digital. Upon entering the transmitter, analog data will first be digitized. In the process of digitization, continuous time may be reduced to discrete time by the process of sampling a source waveform of finite bandwidth. Then the datastream is processed by a source encoder, whose purpose is to represent the source data compactly by a stream of bits called the *source codestream*. At this point, the source data has been reduced to a commonplace stream of bits, superficially displaying no trace of the origin of the source data. Indeed, data from several completely different kinds of sources now may be merged into a single bit stream. The source data might then be encrypted to prevent eavesdropping by an unauthorized receiver. Again, the encrypted bit stream is another commonplace bit stream superficially displaying no trace of its origin.

The datastream is next processed by the channel encoder, which transforms the datastream into a new datastream called the *channel codestream*. The channel codestream has redundancy in the form of elaborate cross checks built into it, so that errors arising in the channel can be corrected by using the cross checks. Redundancy may also be added to remove subpatterns of data that are troublesome to transmit. The symbols of the new datastream might be binary or might be symbols from a larger alphabet called the *channel alphabet*. The stream of channel codewords is passed to the modulator, which converts the sequence of discrete code symbols into a continuous function of time called the *channel waveform*. The modulator does this by replacing each symbol of the channel codeword by the corresponding analog symbol from a finite set of analog symbols composing the *modulation alphabet*. Here, discrete time is also reconverted to continuous time by some form of interpolation. The sequence of analog symbols composes the channel waveform, which is either transmitted through the channel directly or after it is further modified to spread its bandwidth. The reason that bandspreading might be used is to protect the signal from some kinds of fading or interference, possibly intentional interference created by an adversary. The input to the channel is the channel waveform formed by the transmitter. The channel waveform is now a continuous-time waveform. Although the source waveform might also have been a continuous-time waveform, the appearance and properties of the channel waveform will be quite different from those of the source waveform. The channel waveform then passes through the channel where it may be severely attenuated and usually changed in other ways.

The input to the receiver is the output of the channel. Because the channel is subject to various types of noise, dispersion, distortion, and interference, the waveform seen at the channel output differs from the waveform at the channel input. The waveform will always be subjected to thermal noise in the receiver, which is additive gaussian noise, and this is the disturbance that we shall study most thoroughly. The waveform may also be subjected to many kinds of impulsive noise, burst noise, or other forms of nongaussian noise. Upon entering the receiver, if the waveform has been bandspread in the transmitter, it is first despread. The demodulator may then convert the received waveform into a stream of discrete channel symbols based on a best estimate of each transmitted symbol. Sometimes the demodulator makes errors because the received waveform has been impaired and is not the same as the waveform that was transmitted. Perhaps to quantify its confidence, the demodulator may append confidence annotations to each demodulated symbol. The final sequence of symbols from the demodulator is called the *received word* or the *senseword*. It is called a *soft senseword* if it is not reduced to the channel input alphabet or if it includes confidence annotations. The symbols of the senseword need not match those of the transmitted channel codeword; the senseword symbols may take values in a different alphabet.

The function of the channel decoder is to use the redundancy in the channel codeword to correct the errors in the senseword and then to produce an estimate of the datastream that appeared at the input to the channel encoder. If the datastream has been encrypted, it is now decrypted to produce an estimate of the sequence of source codewords. Possibly at this point the datastream contains source codewords from more than one source, and these source codewords must be demultiplexed. If all errors have been corrected by the channel decoder, each estimated source codeword matches the original source codeword. The source decoder performs the inverse operation of the source encoder and delivers its output to the user.

The modulator and the demodulator are studied under the term *modulation theory*. Modulation theory is the core of digital communications and the subject of this book. The methods of baseband modulation and baseband demodulation are studied in Chapters 2 and 3, while passband modulation and passband demodulation are studied in Chapter 5 and Chapter 6. The formal justification for the structure of the optimal receiver will be given with the aid of the maximum-likelihood principle, which is studied in Chapter 7.

The receiver also includes other functions, such as equalization and synchronization, shown in Figure 1.3, that are needed to support demodulation. The corresponding structure of the optimal receiver will not fully emerge until these functions are developed as a consequence of the maximum-likelihood principle in Chapter 7. The receiver may also include intentional nonlinearities, perhaps meant to clip strong interfering signals, or to control dynamic range. These are studied in Chapter 11.

1.4 Networks for digital communication

Digital point-to-point communication systems can be combined to form digital communication networks. Modern communication networks were developed relatively recently, so they are mostly digital. The main exception to this is the telephone network which began early as an analog network, and was converted piecemeal to a digital network through the years, but still with some vestiges of analog communication. The telephone network is a kind of network in which switching (*circuit switching*) is used to create a temporary communication channel (a *virtual channel*) between two points. A broadcast system, shown in Figure 1.4, might also be classified as a kind of communication network. However, if the same waveform is sent to all receivers, the design of a broadcast system is really quite similar to the design of a point-to-point communication system. Therefore both the early telephone network and a broadcast system, in their origins, can be seen as kinds of analog point-to-point communication systems. Other, more recent, communication networks are digital.

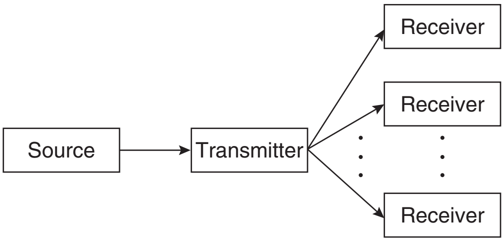

Figure 1.4. Broadcast communications.