Communications Channels

(Based on Blahut's Digital Transmission of Information)

The transfer of messages between individuals or between cells is an activity that is fundamental to all forms of life, and in biological systems, delicate mechanisms have evolved to support this message transfer. Each of these messages traverses an environment capable of producing various kinds of distortion in the message, but the message must be understood despite this distortion. Yet, the method of message transmission must not be so complex as to unduly exhaust the resources of the transmitter or the receiver.

Man-made communication systems also transfer data through a noisy environment. This environment is called a communication channel. A very important channel is the one in which a message is represented by an electromagnetic wave that is sent to a receiver, where it appears contaminated by noise. The transmitted messages must be protected against distortion and noise in the channel. The first such communication systems protected their messages from the environment by the simple expedient of transmitting high power. Later, message design techniques were introduced that led to the development of far more sophisticated communication systems. Modern message design is the art of piecing together a number of waveform ideas to meet a set of requirements: generally, to transmit as many bits as is practical within the available power and bandwidth. It is by the performance at low transmitted power that one judges the quality of a digital communication system. The purpose of this course is to develop in a rigorous mathematical setting the modem waveform techniques for the digital transmission of information.

1.1 TRANSMISSION OF INFORMATION

Communication theory studies methods for the design of messages for the transmission of information. As such, it is a branch of the subject of information theory. Communication theory is that part of information theory that is concerned with the explicit design of systems to convey messages and with the performance of those systems. Digital communication theory is that part of communication theory in which digital techniques play a prominent role in the communication process, either because the information to be transmitted is digital or because the information is temporarily represented in digital form for the purpose of transmission.

An overview of a digital communication system is shown in Fig. 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 a 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. At a fundamental level, a communication channel is normally an analog channel in that it transmits waveforms. Digital data must be modulated into an analog waveform for passage through the channel. The source data may arise as digital data or as

analog data. The user may require the data be delivered in digital form or in analog form. When the source data is analog, perhaps a continuous- time analog waveform, a digital communication system converts it to digital data, processes it, then converts it to a continuous-time analog waveform for passage through the channel. By means of sampling and quantization, the analog waveform out of the channel is converted to digital data for processing even though it may ultimately be required in analog form by the user. This constant conversion between digital data and analog data may seem inefficient but,

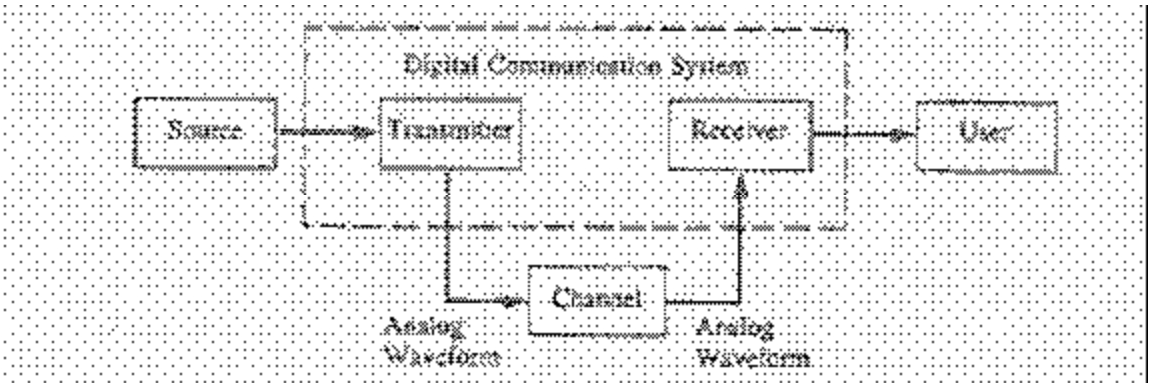

Figure 1.1 Overview of a digital communication system.

because so much can be done to the data while it is in digital form, it is worthwhile. Besides, the analog waveform passing through the channel may be completely different from the analog waveform generated by the source.

In contrast to digital modulation is analog modulation, which today is widely, used in radio, television, and phonography. Analog modulation techniques make relatively superficial changes to the signal in order to get 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 much 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 expensive and it is important to make the best use of a channel.

Even applications that might appear to be intrinsically analog applications, such as broadcast television, can be partitioned into two tasks—the task of delivering so many bits per second to the viewer and the task of representing the video signal by the available number of bits per second. This is the important and 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 loss, into the task of forming a binary representation of the source output and the task of sending a binary sequence through the channel. For digital transmission to be effective, both of these tasks must be implemented efficiently

using methods described in this course. Otherwise there will be disadvantages such as increased bandwidth or larger transmitted power. The data must be compressed, then modulated into a transmission waveform, perhaps one that carries multiple bits per second per hertz.

The only disadvantage of digital communication that must be accepted is that it is not always compatible with the existing analog communication receivers of earlier generations; it would be expensive and unpopular to discard existing AM and FM radio receivers in favor of a new high-fidelity digital format. Other disadvantages that are often cited are really not compelling; any validity that they once had has crumbled under the progress of technology. Specifically, cost was once a disadvantage of digital communication systems but no longer is. Bandwidth expansion is often mentioned as a disadvantage of a simple system, but with modern data compaction, and multilevel signaling, one can actually reduce the bandwidth by using digital communication.

Another disadvantage 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 received by 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 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 probability of bit error. It makes no difference to the link whether the bits represent digitized voice or computer programs. Many kinds of data source can share a common digital communication link, and the many kinds of communication link are suitable for every data source. A binary data stream 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 data stream is compatible with standardized encryption and antijam equipment. Errors due to noise and interference are almost completely suppressed by error correction codes.

The digital data stream can be regenerated at every repeater that it passes through so that the effect of additive noise does not accumulate in the signal. Analog repeaters, on the other hand, consist of amplifiers that amplify both signal and noise. Thus, noise accumulates as an analog signal passes through a series of repeaters.

Finally, note that digital communication systems are built in large part from digital circuitry. A digital data format is compatible with the many other digital subsystems in a large system. Data can be readily buffered in random-access digital memories or on magnetic discs. Many functions of a modulator/ demodulator can be programmed into a microprocessor or designed into a special-purpose digital integrated circuit.

1.2 A BRIEF HISTORICAL SURVEY

It is a curiosity of technological history that the earliest communication systems such as telegraphy (1866) were actually digital communication systems. The time-honored Morse code is a waveform for digital communication in that it uses a discrete alphabet. Telegraphy created a communications industry but lacked the popular appeal of later

analog communication systems such as the telephone (1876) and the phonograph (1877). The analog systems have been dominant for most of the twentieth century.

The earliest broadcast systems for communication were concerned with the transfer of analog continuous signals, first radio (1920) and then television signals. 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, Indications can be seen, however, that analog modulation is becoming outdated even for those applications, and only the enormous investment in existing equipment will forestall the inevitable. 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 strength in the 1970s. Even the analog phonograph record, after 100 years, has been superseded by the compact disk.

The earliest radio transmitters used amplitude modulation. This form of analog modulation maps a signal *s(t)* into a waveform

$$c(t) = [1 + ms(t)]\cos 2pf_0 t$$

where *m* is a small constant called the modulation index such that *ms(t)* << 1, and where *m* is a constant called the carrier frequency such that f0 is large compared to the largest frequency for which *S(f)*, the Fourier transform of *s(t)*, is nonzero. 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. Therefore, amplitude modulation became very popular early on even though the fidelity of the received signal is not noteworthy.

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

$$c(t) = \sin\left(2pf_0 t + \int_0^t ms(x)dx\right)$$

where, again, the carrier frequency f0 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 argument was that the term *ms(t)* is an "instantaneous frequency" perturbing the carrier frequency f0 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 (L922) argued that this is an ill-considered plan, as is easily seen by looking at the approximation.

$$c(t) \approx \sin 2pf_0 t + \cos 2pf_0 t\left[m\int_0^t s(x)dx\right]$$

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) rewakened 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 amplitude modulation and frequency 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 in use to match the analog source signal more closely to the noise characteristics of the channel. All of the methods for modifying the techniques of analog modulation are stopgap methods. They do not attack the deficiencies of analog modulation head on. Eventually, such methods will be abandoned for the most part in favor of digital modulation.

The theory of communication became much more mathematical when it became widely appreciated that communication is intrinsically a statistical process. There is no point in transmitting a predetermined message if the receiver already knows the message. 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; randomness is an essential ingredient in communication because of the statistical nature of messages. Randomness is also an essential ingredient in communication because of noise in the channel. This statistical view of communication encompassing both random messages and noisy channels was popularized 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 counterintuitive view that the waveform could be designed so as to make the channel noise essentially inconsequential in the quality of the received waveform. Shannon realized that combating noise was a job for both the transmitter and receiver, not for the receiver alone.

In his papers, Shannon laid a firm foundation for the development of digital communication, A somewhat more applied paper, transitional between analog and digital communication, is due to Oliver, Pierce, and Shannon (1948), This period seems to be the time when people were first thinking deeply about the fundamental problems and the return to digital signaling began, but there were many studies and applications of digital signaling earlier as in the work of Nyquist (1924, 1928) and Hartley (1928). Aschoff (1983) gives a good history of digital signaling.

There is a misconception that is sometimes cited as a disadvantage of digital modulation; digital modulation of an analog signal is said to require larger bandwidth than direct analog modu1aion, This is not true. It was true when digital communication waveforms were limited to simple pulse code modulation. Today by using data compression, data compaction, and trellis-coded modulation that conveys multiple bits per second per hertz, one can have a digital communication system using less bandwidth than an analog communication system, and the fidelity of such systems is rapidly

improving. The superiority of digital communication is more than just an empirical observation of recent engineering trends; its optimality is implicit in Shannon's original theorems.

1.3 POINT-TO-POINT DIGITAL COMMUNICATION

A simple block diagram of a point-to-point digital communication system was shown in Fig. 1.1. Even though we will use the terminology of communication theory to describe it, the model in Fig. 1.1 is quite general and can be applied to a variety of situations. One can interpret many other information-handling systems, such as mass storage systems, in terms of this model. It only is necessary to translate the terminology and to identify the properties of the channel. The identification of the channel may be somewhat arbitrary because some of the physical components such as amplifiers might sometimes be considered to be part of the channel and might sometimes be considered as part of the modulator and demodulator. The boxes labeled channel, data source, and data user in Fig. 1.1 are those parts of the system that are not under the control of the designer.

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 device is designed so it can play either the role of the transmitter or the role of the receiver (or of both simultaneously); then the modulator and demodulator can be combined into a box called the modem. The term "modem" might also be construed to include the encoder and decoder as well, and might also include other functions that extend beyond modulation and demodulation. We prefer to use the terms "transmitter" and "receiver" as the broader terms that include supporting functions. Figures 1.2 and 1.3 show the functions normally included in the transmitter and receiver.

The trend 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 source data, when reduced to a stream of bits, might be transmitted through any one of several possible channels. To do this, the transmitter and receiver are broken into more detailed functions as described by the block diagrams of Figs. 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, there is no loss in generality if the interface between the

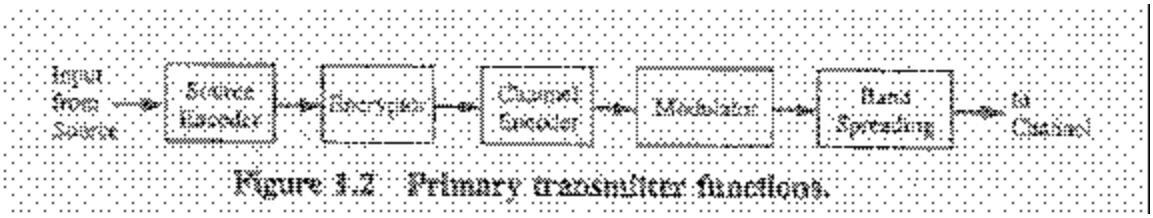

Figure 1.2 Primary transmitter functions.

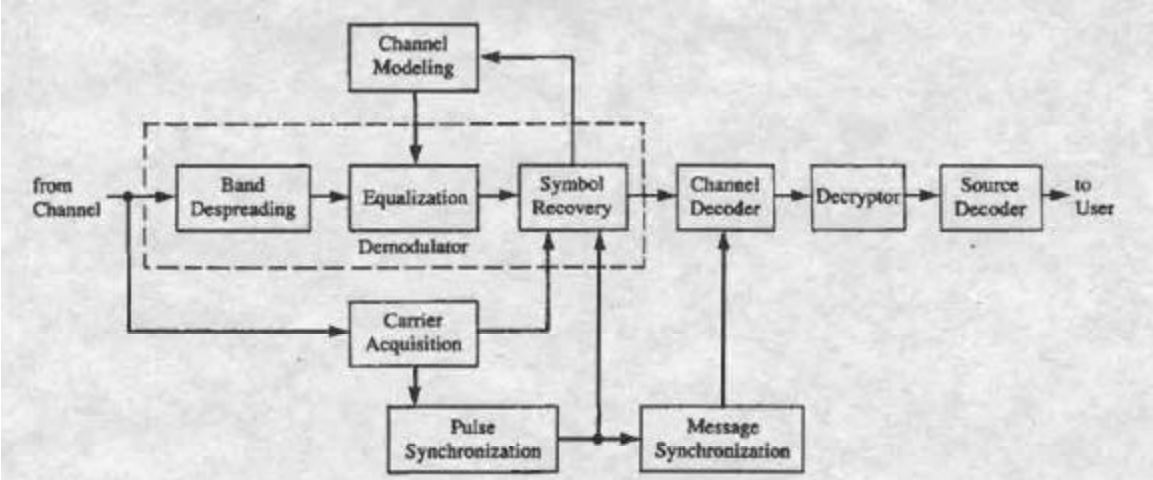

Figure 1.3 Primary receiver functions.

source encoder and channel encoder and the interface between the channel decoder and source decoder are specified to be binary bit streams.

The source data may be analog or digital. Upon entering the transmitter, analog data must first be digitized. In the process of digitization, continuous time may be reduced to discrete time by the process of sampling of a source waveform of finite bandwidth. Then the data stream is processed by a source encoder, whose purpose is to represent the source data compactly by a stream of bits called the source codewords. 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 may 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 bits are next processed by the channel encoder, which transforms the bit stream into a new data' stream called the channel codestream, The new data stream has more redundancy than the stream of source codewords; the purpose of the redundancy is to match the constraints of the channel and so that errors arising in the channel can be corrected. The symbols of the new data stream could be binary or could be symbols from a larger alphabet called the channel alphabet. The stream of channel codewords is passel 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 reconverted to continuous time. The sequence of analog symbols composes the channel waveform, which is transmitted through the channel directly or after the bandwidth is intentionally spread. The reason bandspreading might be used is to protect the signal from some kinds of interference, possibly intentional interference created by an adversary.

The input to the channel is the output of the transmitter. The channel waveform, as might be the source waveform, is now a continuous-time waveform. However, it will have an appearance and properties quite different from the source waveform that appeared at the input to the transmitter.

The input to the receiver is the output of the channel. Because the channel is subject to various types of noise, 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 oil a best estimate of each transmitted symbol. Sometimes it makes errors because the received waveform is not the same as was transmitted. Perhaps to quantify its confidence, the demodulator may append additional data to each demodulated symbol. The demodulated sequence of symbols, possibly including confidence annotations, is called the received word. The symbols of the received word do not always match those of the channel codeword. The function of the channel decoder is to use the redundancy in a channel codeword to correct the errors in the received word, and then to produce an estimate of the bit stream that appeared at the input to the channel encoder. If the bit stream has been encrypted, it is now decrypted to produce an estimate of the sequence of source codewords. Possibly at this point the bit stream contains source codewords from more than one source, and these source codewords must be deinterleaved. 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 channel encoder and channel decoder are commonly split into two functions: that of implementing error control to negate the effects of channel noise; and that of preparing the sequence of transmitted symbols to be compatible with channel constraints. These functions are studied under the terms data transmission and data translation (or under the terms error-control coding and constrained channel coding). Data transmission encompasses the various discrete methods, such as the use of error-correcting codes, for combating noise and errors in the communication channel. Data translation encompasses various methods for preventing troublesome symbol patterns from appearing at the input to the modulator.

The modulator and the demodulator are studied under modulation theory. This is the core of digital communications. The main elements of the structure of the optimal receiver will be found with the aid of the maximum- likelihood principle.

The source encoder and source decoder are commonly split into two functions: that of expressing the output of the source compactly, and that of abridging the output of the source. We shall study these under the terms data compaction and data compression. Data compaction encompasses various methods for reducing redundancy in a stream of digital data. Data compression encompasses the various methods for reducing the entropy of a stream of data. This includes primarily methods such as quantization for digitizing analog signals. There are numerous parallels between codes for data compression and codes for data transmission. There are also numerous parallels between sampling theory, used to turn a continuous-time waveform into a discrete-time sequence, and modulation theory, used to turn a digital sequence into a continuous-time modulation waveform.

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

The functional anatomy of a digital communication system that we have sketched hem also serves as an outline for designing a communications system. The potency of the ideas that emerge may be underscored by mentioning how advantages may come froth unexpected directions. For example, the purpose of an error-control code may at first appear to be to improve the quality of the data, However, a subtler view is equally valid and coming into vogue. The error-control code can be thought of as a way to reduce transmitted power and antenna size in a modem digital communication system, this because the use of an error-control code allows the system to run at lower signal-to-noise ratio.

1.4 NETWORKS FOR DIGITAL COMMUNICATION

In contrast to digital point-to-point communication systems are digital communication networks. Communication networks became important relatively recently, so they are mostly digital. The main exception to this is the telephone network. The telephone network is a kind of network in which switching (circuit switching) is used to create a temporary point-to-point communication channel (a virtual channel). A broadcast system, shown in Fig. 1.4, might also be classified as a communication network. However, if the same waveform is sent to all receivers, the design of a broadcast system is virtually the same as the design of a point-to-point communication system. Therefore, both the telephone network and a broadcast system can be seen as kinds of analog point-to-point communication systems. Most other communication networks are digital.

Figure 1.5 shows a modem form of a communication network, the multi- access communication network. It consists of a single channel, to which is

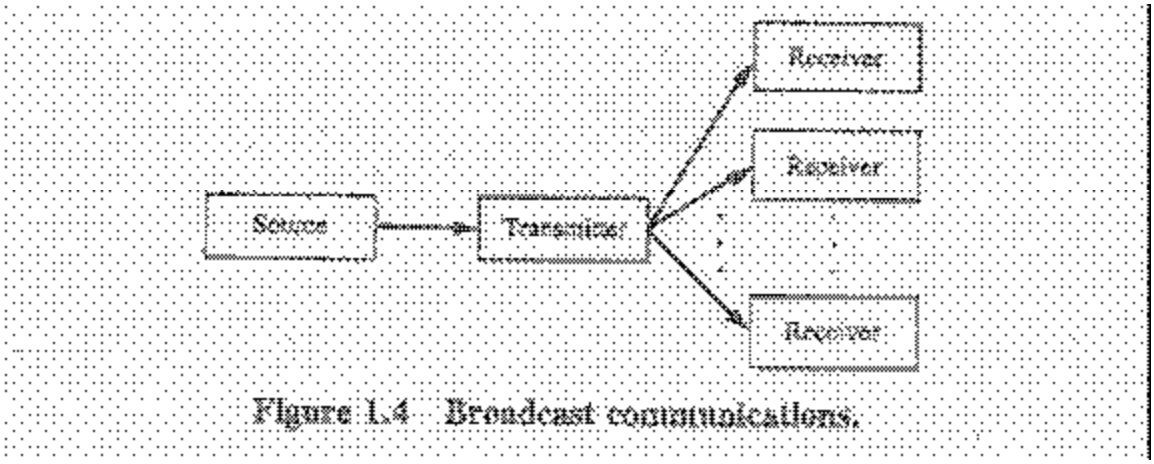

Figure 1.4 Broadcast communications.

attached a multitude of users who exchange messages. The rules by which the users take turns using the channel are called channel protocols. The protocol is the new element that makes a digital communication network more complicated than a digital point-to-point communication system.

The simplest multiaccess channel is the fixed-assignment multiaccess channel, of which lime-division multiaccess signaling is a good example. Each user is preassigned a fixed time interval called a time slot during which it can transmit, One disadvantage of the time-division protocol is that it is inefficient; if a user has nothing to transmit, its time slot is wasted. Another disadvantage is that the set of users needs to be fairly well specified so that time slots can be assigned. Each user, both transmitter and receiver, needs to know its slot assignment and needs to have its clock synchronized to system time.

At the other extreme are networks that take the form of a digital version of citizen's band radio. The set of users is very large, not very well defined, and each individual user transmits only rarely but at random times and at its own convenience. Then the channel is known as a demand-assignment multiaccess channel. One kind of protocol for this case is the kind known as a contention resolution algorithm. Users transmit whenever it suits them but, in the event of a conflict, all users but one stop and retransmit at a later tithe. Algorithms to resolve contention are in wide use but are still incompletely understood. From a practical point of view one wishes to be ensured that a contention resolution algorithm will successfully deal with every possible situation of contention; otherwise the channel might fail because the protocol has deadlocked.

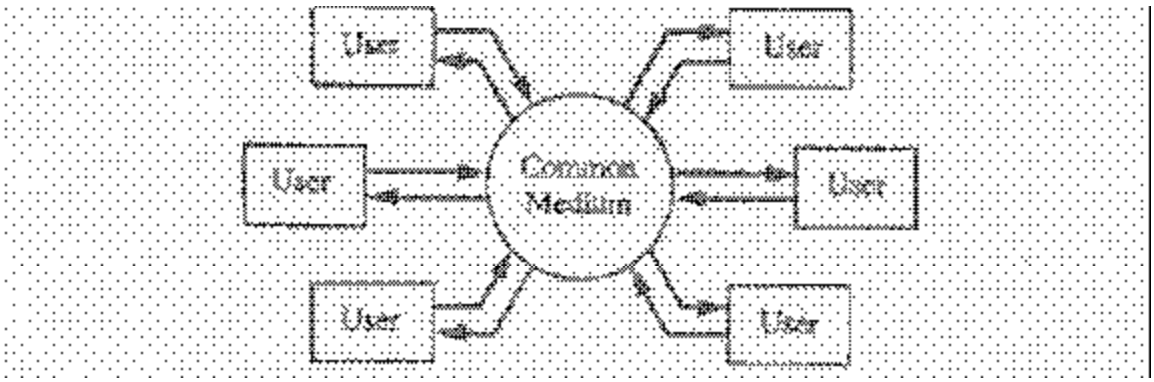

Figure 1.5 A multiaccess communication network.