AN INTRODUCTION TO ANALOG AND DIGITAL COMMUNICATIONS

Simon Haykin

Library of Congress Cataloging in Publication Data:

Haykin, Simon S., 1931–
An introduction to analog and digital communications/Simon Haykin
p. cm.
References and Bibliography: pp. 635–639
Includes index.
ISBN 0-471-85978-8
1. Telecommunication systems. I. Title.
TK5101.H373 1989
821.38—dc 18

88-15512
CIP

Printed in the United States of America

10 9 8 7 6 5 4 3 2

CHAPTER ONE

INTRODUCTION

1.1 COMMUNICATIONS

Communication enters our daily lives in so many different ways that it is easy to overlook the multitude of its facets. The telephones in our homes and offices make it possible for us to communicate with others, no matter how far away. The radio and television sets in our living rooms bring us entertainment from near as well as far-away places. Communication by radio or satellite provides the means for ships on the high seas, aircraft in flight, and rockets and exploratory probes in space to maintain contact with their home bases. Communication keeps the weather forecaster informed of atmospheric conditions that are

measured by a multitude of sensors. Communication makes it possible for computers to interact. The list of applications involving the use of communications in one way or another goes on.[1]

In the most fundamental sense, *communication* involves implicitly the transmission of information from one point to another through a succession of processes, as described here:

1. The generation of a thought pattern or image in the mind of an originator.
2. The description of that thought pattern or image, with a certain measure of precision, by a set of aural or visual symbols.
3. The encoding of these symbols in a form that is suitable for transmission over a physical medium (channel) of interest.
4. The transmission of the encoded symbols to the desired destination.
5. The decoding and reproduction of the initial symbols.
6. The re-creation of the original thought pattern or image—with a definable degradation in quality—in the mind of a recipient, with the degradation being caused by imperfections in the system.

The form of communication just described involves a thought pattern or image originating in a human mind. Of course, there are many other forms of communication that do not directly involve the human mind in real time. In space exploration, for example, human decisions may enter only the commands sent to the space probe or to the computer responsible for processing images of far-away planets (e.g., Mars, Jupiter, Saturn) that are sent back by the probe. In computer communications, human decisions enter only in setting up the computer programs or in monitoring the results of computer processing.

Whatever form of communication is used, some basic signal-processing operations are involved in the transmission of information. The next section describes the different types of signals encountered in the study of communication systems. The signal-processing operations of interest are highlighted later in the chapter.

1.2 SIGNALS AND THEIR CLASSIFICATIONS

For our purposes, a *signal* is defined as a single-valued function of time that conveys information. Consequently, for every instant of time there is a unique value of the function. This value may be a real number, in which case we have a *real-valued signal,* or it may be a complex number, in which case we have a *complex-valued signal.* In either case, the independent variable (namely, time) is real-valued.

[1]For an essay on communications, see Berkner (1962).

For a given situation, the most useful method of signal representation hinges on the particular type of signal being considered. Depending on the feature of interest, we may identify four different methods of dividing signals into two classes:

1. PERIODIC SIGNALS, NONPERIODIC SIGNALS

A *periodic signal* $g(t)$ is a function that satisfies the condition

$$g(t) = g(t + T_0) \tag{1.1}$$

for all t, where t denotes time and T_0 is a constant. The smallest value of T_0 that satisfies this condition is called the *period* of $g(t)$. Accordingly, the period T_0 defines the duration of one complete cycle of $g(t)$.

Any signal for which there is no value of T_0 to satisfy the condition of Eq. 1.1 is called a *nonperiodic* or *aperiodic signal.*

2. DETERMINISTIC SIGNALS, RANDOM SIGNALS

A *deterministic signal* is a signal about which there is no uncertainty with respect to its value at any time. Accordingly, we find that deterministic signals may be modeled as completely specified functions of time.

On the other hand, a *random signal* is a signal about which there is uncertainty before its actual occurrence. Such a signal may be viewed as belonging to an ensemble of signals, with each signal in the collection having a different waveform. Moreover, each signal within the ensemble has a certain *probability* of occurrence.

3. ENERGY SIGNALS, POWER SIGNALS

In communication systems, a signal may represent a voltage or a current. Consider a voltage $v(t)$ developed across a resistor R, producing a current $i(t)$. The *instantaneous power* dissipated in this resistor is defined by

$$p = \frac{v^2(t)}{R} \tag{1.2}$$

or, equivalently,

$$p = Ri^2(t) \tag{1.3}$$

In both cases, the instantaneous power p is proportional to the squared amplitude of the signal. Furthermore, for a resistance R of 1 ohm, we see that Eqs. 1.2 and 1.3 take on the same mathematical form. Accordingly, in signal analysis it is customary to work with a 1-ohm resistor, so that,

regardless of whether a given signal $g(t)$ represents a voltage or a current, we may express the instantaneous power associated with the signal as

$$p = g^2(t) \tag{1.4}$$

Based on this convention, we define the *total energy* of a signal $g(t)$ as

$$\begin{aligned} E &= \lim_{T\to\infty} \int_{-T}^{T} g^2(t)dt \\ &= \int_{-\infty}^{\infty} g^2(t)dt \end{aligned} \tag{1.5}$$

and its *average power* as

$$P = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} g^2(t)dt \tag{1.6}$$

We say that the signal $g(t)$ is an *energy signal* if and only if the total energy of the signal satisfies the condition

$$0 < E < \infty$$

We say that the signal $g(t)$ is a *power signal* if and only if the average power of the signal satisfies the condition

$$0 < P < \infty$$

The energy and power classifications of signals are mutually exclusive. In particular, an energy signal has zero average power, whereas a power signal has infinite energy. Also, it is of interest to note that, usually, periodic signals and random signals are power signals, whereas signals that are both deterministic and nonperiodic are energy signals.

4. ANALOG SIGNALS, DIGITAL SIGNALS

An *analog signal* is a signal with an *amplitude* (i.e., value of the signal at some fixed time) that varies continuously for all time; that is, *both amplitude and time are continuous over their respective intervals*. Analog signals arise when a physical waveform such as an acoustic wave or a light wave is converted into an electrical signal. The conversion is effected by means of a *transducer;* examples include the microphone, which converts sound pressure variations into corresponding voltage or current variations, and the photodetector cell, which does the same for light-intensity variations.

On the other hand, a *discrete-time signal* is defined only at discrete instants of time. Thus, in this case, the independent variable takes on only

discrete values, which are usually uniformly spaced. Consequently, discrete-time signals are described as sequences of samples that may take on a continuum of values. When each sample of a discrete-time signal is *quantized* (i.e., it is only allowed to take on a finite set of discrete values) and then *coded,* the resulting signal is referred to as a *digital signal.* The output of a digital computer is an example of a digital signal. Naturally, an analog signal may be converted into digital form by *sampling in time, then quantizing and coding.*

1.3 FOURIER ANALYSIS OF SIGNALS AND SYSTEMS

In theory, there are many possible methods for the representation of signals. In practice, however, we find that *Fourier analysis,* involving the resolution of signals into *sinusoidal components,* overshadows all other methods in usefulness. Basically, this is a consequence of the well-known fact that the output of a system to a sine-wave input is another sine wave of the same frequency[2] (but with a different phase and amplitude) under two conditions:

1. The system is *linear* in that it obeys the *principle of superposition.* That is, if $y_1(t)$ and $y_2(t)$ denote the responses of a system to the inputs $x_1(t)$ and $x_2(t)$, respectively, the system is linear if the response to the composite input $a_1x_2(t) + a_2x_2(t)$ is equal to $a_1y_1(t) + a_2y_2(t)$, where a_1 and a_2 are arbitrary constants.
2. The system is *time-invariant.* That is, if $y(t)$ is the response of a system to the input $x(t)$, the system is time-invariant if the response to the time-shifted input $x(t - t_0)$ is equal to $y(t - t_0)$, where t_0 is constant.

Given a linear time-invariant system, the *response* of the system to a single-frequency *excitation* represented by the complex exponential time function $A \exp(j2\pi ft)$ is equal to $AH(f) \exp(j2\pi ft)$, where $H(f)$ is the *transfer function* of the system; the complex exponential $\exp(j2\pi ft)$ contains the cosine function $\cos(2\pi ft)$ as its real part and the sine function $\sin(2\pi ft)$ as its imaginary part. Thus, the response of the system exhibits exactly the same variation with time as the excitation applied to the system. This remarkable property of linear time-invariant systems is realized only by using the complex exponential time function.

In the study of communication systems, we are usually interested in a *band of frequencies.* For example, although the average voice spectrum extends well beyond 10 kHz, most of the energy is concentrated in the range of 100 to 600 Hz, and a voice signal lying inside the band from 300 to 3400 Hz gives good articulation. Accordingly, we find that telephone

[2]For a historical account of the concept of frequency, see Manley (1982).

circuits that respond well to the band of frequencies from 300 to 3400 Hz give satisfactory commercial telephone service.

To talk meaningfully about the *frequency-domain description* or *spectrum* of a signal, we need to know the *amplitude* and *phase* of each frequency component contained in the signal. We get this information by performing a Fourier analysis on the signal. However, there are several methods of Fourier analysis available for the representation of signals. The particular version that is used in practice depends on the type of signal being considered. For example, if the signal is periodic, then the logical choice is to use the *Fourier series* to represent the signal as a set of harmonically related sine waves. On the other hand, if the signal is an energy signal, then it is customary to use the *Fourier transform* to represent the signal. Irrespective of the type of signal being considered, Fourier methods are invertible. Specifically, if we are given the complete spectrum of a signal, then the original signal (as a function of time) can be reconstructed exactly. The Fourier analysis of signals and systems is considered in Chapters 2 through 4.

1.4 ELEMENTS OF A COMMUNICATION SYSTEM

The purpose of a *communication system* is to transmit information-bearing signals from a *source,* located at one point, to a *user destination,* located at another point some distance away. When the message produced by the source is not electrical in nature, which is often the case, an input transducer is used to convert it into a time-varying electrical signal called the *message signal.* By using another transducer connected to the output end of the system, a "distorted" version of the message is re-created in its original form, so that it is suitable for delivery to the user destination. The distortion mentioned here is due to inherent limitations in the communication system.

Figure 1.1 is a block diagram of a communication system consisting of three basic components: transmitter, channel, and receiver. The *transmitter* has the function of processing the message signal into a form suitable for

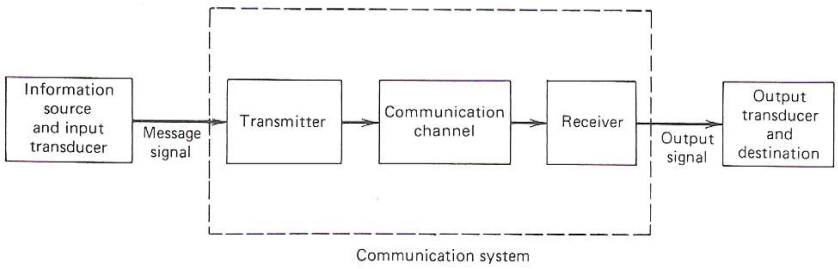

Figure 1.1
Elements of an electrical communication system.

transmission over the channel; such an operation is called *modulation*. The function of the *channel* is to provide a physical connection between the transmitter output and the receiver input. The function of the *receiver* is to process the *received signal* so as to produce an "estimate" of the original message signal; this second operation is called *detection* or *demodulation*.

There are two types of channels, namely, point-to-point channels and broadcast channels. Examples of *point-to-point channels* include wire lines, microwave links, and optical fibers. *Wire lines* operate by guided electromagnetic waves; they are used for local telephone transmission. In *microwave links,* the transmitted signal is radiated as an electromagnetic wave in free space; microwave links are used in long-distance telephone transmission. An *optical fiber* is a low-loss, well-controlled, guided optical medium; optical fibers are used in optical communications.[3] Although these three channels operate differently, they all provide a physical medium for the transmission of signals from one point to another point; hence, the term "point-to-point channels."

Broadcast channels, on the other hand, provide a capability where many receiving stations may be reached simultaneously from a single transmitter. An example of a broadcast channel is a *satellite in geostationary orbit,* which covers about one third of the earth's surface. Thus, three such satellites provide a complete coverage of the earth's surface, except for the polar regions.

1.5 TRANSMISSION OF MESSAGE SIGNALS

To transmit a message (information-bearing) signal over a communication channel, we may use *analog* or *digital* methods. The use of digital methods offers several important operational advantages over analog methods, which include the following:

1. Increased *immunity* to channel noise and external interference.
2. *Flexible operation* of the system.
3. A *common format* for the transmission of different kinds of message signals (e.g., voice signals, video signals, computer data).
4. Improved *security* of communication through the use of encryption.

These advantages are attained, however, at the cost of *increased transmission (channel) bandwidth and increased system complexity*. The first requirement is catered to by the availability of *wideband communication channels* (e.g., optical fibers, satellite channels). The second requirement is taken care of by the use of *very large-scale integration* (VLSI) technology, which offers a cost-effective way of building hardware. Accordingly, there

[3]For a discussion of electronic and photonic (optical) communication systems, see Williams (1987).

is an ever-increasing trend toward the use of digital communications and away from analog communications. This trend is being accelerated by the pervasive influence of digital computers in so many facets of our daily lives. Nevertheless, analog communications remain a force to be reckoned with. Most of the broadcasting systems and a large part of the telephone networks in use today are analog in nature and, moreover, they will remain in service for some time yet. It is therefore important that we understand the operations and requirements of both analog and digital communications.

Notable among the digital methods that may be used for the transmission of message signals over a communication channel is *pulse-code modulation* (PCM). In PCM, the message signal is *sampled, quantized,* and then *encoded.* The sampling operation permits representation of the message signal by a sequence of samples taken at uniformly spaced instants of time. Quantization trims the amplitude of each sample to the nearest value selected from a finite set of representation levels. The combination of sampling and quantization permits the use of a *code* (e.g., binary code) for the transmission of a message signal. Pulse-code modulation and related methods of *analog-to-digital conversion* are covered in Chapter 5.

When digital data are transmitted over a band-limited channel, a form of interference known as *intersymbol interference* may result. The effect of intersymbol interference, if left uncontrolled, is to severely limit the rate at which digital data may be transmitted over the channel. The cure for controlling the effects of intersymbol interference lies in *shaping* the transmitted pulse representing a binary symbol 1 or 0. Intersymbol interference is considered in Chapter 6.

To transmit a message signal (be it in analog or digital form) over a *band-pass* communication channel (e.g., telephone channel, microwave radio link, satellite channel) we need to modify the message signal into a form suitable for *efficient transmission* over the channel. Modification of the message signal is achieved by means of a process known as *modulation.* This process involves varying some parameter of a carrier wave in accordance with the message signal in such a way that the spectrum of the modulated wave matches the assigned channel bandwidth. Correspondingly, the receiver is required to re-create the original message signal from a degraded version of the transmitted signal after propagation through the channel. The re-creation is accomplished by using a process known as *demodulation,* which is the inverse of the modulation process used in the transmitter.

There are other reasons for performing modulation. In particular, the use of modulation permits *multiplexing,* that is, the simultaneous transmission of signals from several message sources over a common channel. Also, modulation may be used to convert the message signal into a form less susceptible to noise and interference.

A carrier wave commonly used to perform modulation is the sinusoidal wave. Such a carrier wave has three independent parameters that can be varied in accordance with the message signal; they are the carrier ampli-

tude, phase, and frequency. The corresponding forms of modulation are known as *amplitude modulation, phase modulation,* and *frequency modulation,* respectively. Amplitude modulation offers simplicity of implementation and a transmission bandwidth requirement equal to twice the message bandwidth; the *message bandwidth* is defined as the extent of significant frequencies contained in the message signal. With special processing, the transmission bandwidth requirement may be reduced to a value equal to the message bandwidth, which is the minimum possible. Phase and frequency modulation, on the other hand, are more complex, requiring transmission bandwidths greater than that of amplitude modulation. In exchange, they offer a superior noise immunity, compared to amplitude modulation. Modulation techniques for analog and digital forms of message signals are studied in Chapter 7.

1.6 LIMITATIONS AND RESOURCES OF COMMUNICATION SYSTEMS

Typically, in propagating through a channel, the transmitted signal is distorted because of *nonlinearities* and *imperfections in the frequency response of the channel.* Other sources of degradation are *noise* and *interference* picked up by the signal during the course of transmission through the channel. Noise and distortion constitute two basic *limitations* in the design of communication systems.

There are various sources of noise, internal as well as external to the system. Although noise is random in nature, it may be described in terms of its *statistical properties* such as the *average power* or the spectral distribution of the average power. The mathematical discipline that deals with the statistical characteristics of noise and other random signals is *probability theory.* A discussion of probability theory and the related subject of *random processes* is presented in Chapter 8. Sources of noise and related system calculations are covered in Appendix C.

In any communication system, there are two primary communication resources to be employed, namely, *average transmitted power* and *channel bandwidth.* The average transmitted power is the average power of the transmitted signal. The channel bandwidth defines the range of frequencies that the channel can handle for the transmission of signals with satisfactory fidelity. A general system design objective is to use these two resources as efficiently as possible. In most channels, one resource may be considered more important than the other. Hence, we may also classify communication channels as *power-limited* or *band-limited.* For example, the telephone circuit is a typical band-limited channel, whereas a deep-space communication link or a satellite channel is typically power-limited.

The transmitted power is important because, for a receiver of prescribed *noise figure,* it determines the allowable separation between the transmitter and receiver. Stated in another way, for a receiver of prescribed noise figure and a prescribed distance between it and the transmitter, the available transmitted power determines the *signal-to-noise ratio* at the receiver

input. This, in turn, determines the *noise performance* of the receiver. Unless this performance exceeds a certain design level, the transmission of message signals over the channel is not considered to be satisfactory.

The effects of noise in analog communications are evaluated in Chapter 9. This evaluation is traditionally done in terms of signal-to-noise ratios. In the case of digital communications, however, the preferred method of assessing the noise performance of a receiver is in terms of the *average probability of symbol error.* Such an approach leads to considerations of optimum receiver design. In this context, the *matched filter* offers optimum performance for the detection of pulses in an idealized form of receiver (channel) noise known as *additive white Gaussian noise.* As such, the matched-filter receiver or its equivalent, the *correlation receiver,* plays a key role in the design of digital communication systems. The matched filter and related issues are studied in Chapter 10.

Turning next to the other primary communication resource, channel bandwidth, it is important because, for a prescribed band of frequencies characterizing a message signal, the channel bandwidth determines the number of such message signals that can be *multiplexed* over the channel. Stated in another way, for a prescribed number of independent message signals that have to share a common channel, the channel bandwidth determines the band of frequencies that may be allotted to the transmission of each message signal without discernible distortion.

There is another important role for channel bandwidth, which is not that obvious. Specifically, channel bandwidth and transmitted (signal) power are *exchangeable* in that we may trade off one for the other for a prescribed system performance. The choice of one modulation scheme over another for the transmission of a message signal is often dictated by the nature of this trade-off. Indeed, the interplay between channel bandwidth and signal-to-noise ratio, and the limitation that they impose on communication, is highlighted most vividly by Shannon's famous *channel capacity theorem.*[4] Let B denote the channel bandwidth, and SNR denote the received signal-to-noise ratio. The channel capacity theorem states that ideally these two parameters are related by

$$C = B \log_2(1 + \text{SNR}), \text{ bits/s} \tag{1.7}$$

where C is the *channel capacity,* and a *bit* refers to a *bi*nary digi*t*. The channel capacity is defined as the maximum rate at which information may be transmitted without error through the channel; it is measured in *bits per second.* Equation 1.7 clearly shows that for a prescribed channel capacity, we may reduce the required SNR by increasing the channel band-

[4]In 1948, Shannon published a paper that laid the foundations of communication theory (Shannon, 1948). The channel capacity theorem is one of three theorems presented in that classic paper.

width B. Moreover, it provides an idealized framework for comparing the noise performance of one modulation system against another.

Finally, mention should be made of the issue of *system complexity*. We usually find that the efficient exploitation of channel bandwidth or transmitted power or both is achieved at the expense of increased system complexity. We therefore have to keep the issue of system complexity in mind, alongside that of channel bandwidth and transmitted power when considering the various trade-offs involved in the design of communication systems.