4

INFORMATION AND CHANNEL CAPACITY

4.1 INTRODUCTION

In this chapter, we attempt to answer two basic questions that arise in the analysis and design of communication systems: (1) Given an information source, how do we evaluate the "rate" at which the source is emitting information? (2) Given a noisy communication channel, how do we evaluate the maximum "rate" at which "reliable" information transmission can take place over the channel? We develop answers to these questions based on probabilistic models for information sources and communication channels.

produces

Information sources can be classified into two categories: analog (or continuous-valued) and discrete. Analog information sources, such as a microphone actuated by a voice signal, emit a continuous-amplitude, continuous-time electrical waveform. The output of a discrete information source such as a teletype consists of sequences of letters or symbols. Analog information sources can be transformed into discrete information sources through the process of sampling and quantizing. In the first part of this chapter we will deal with models for discrete information sources as a prelude to our study of digital communication systems.

A discrete information source consists of a discrete set of letters or *alphabet of symbols*. In general, any *message* emitted by the source consists of a string or sequence of symbols. The symbols in the string or sequence are

emitted at *discrete moments*, usually at a fixed time rate, and each symbol emitted is chosen from the source alphabet. Every message coming out of the source contains some information, but some messages convey more information than others. In order to quantify the information content of messages and the average information content of symbols in messages, we will define a measure of information in this chapter. Using the average information content of symbols and the symbol rate, we will then define an average information rate for the source.

If the units of information are taken to be binary digits or bits, then the average information rate represents the minimum average number of bits per second needed to represent the output of the source as a binary sequence. In order to achieve this rate, we need a functional block in the system that will replace strings of symbols by strings of binary digits. We will discuss a procedure for designing this functional block, called the source encoder, in the first part of this chapter.

In the second part of this chapter, we will develop statistical models that adequately represent the basic properties of communication channels. For modeling purposes, we will divide communication channels into two categories: analog channels and discrete channels. An analog channel accepts a continuous-amplitude continuous-time electrical waveform as its input and produces at its output a noisy smeared version of the ~~input waveform. A~~ discrete channel accepts a sequence of symbols as its input and produces an output sequence that is a replica of the input sequence, except for occasional errors. We will first develop models for discrete communication channels and derive the concept of "capacity" of a communication channel. The channel capacity is one of the most important parameters of a data communication system and it represents the maximum rate at which data can be transferred over the channel with an arbitrarily small probability of error.

While expressions for the capacity of discrete channels are easily derived, such is not the case when we deal with the continuous portion of the channel. For this portion we will simply state the model of the channel and then the expression for the capacity of the channel. For the case of a bandlimited channel with bandwidth B, Shannon has shown that the capacity C is equal to C = $B\log_2(1+S/N)$, where S is the average signal power at the output of the channel and N is the average power of the bandlimited Gaussian noise that accompanies the signal. While we will not attempt to derive this expression for channel capacity, we will consider in detail the implications of $C = B\log_2(1+S/N)$ in the design of communication systems.

The material contained in this chapter is based on the pioneering work of Shannon. In 1948, he published a treatise on the mathematical theory of communication in which he established basic theoretical bounds for the performances of communication systems. Shannon's theory is based on probabilistic models for information sources and communication channels. We present here some of the important aspects of his work.

4.2 MEASURE OF INFORMATION

4.2.1 Information Content of a Message

The output of a discrete information source is a message that consists of a sequence of symbols. The actual message that is emitted by the source during a message interval is selected at random from a set of possible messages. The communication system is designed to reproduce at the receiver either exactly or approximately the message emitted by the source.

[illegible]

As mentioned earlier, some messages produced by an information source contain more information than other messages. The question we ask ourselves now is, how can we measure the information content of a message quantitatively so that we can compare the information content of various messages produced by the source? In order to answer this question, let us first review our intuitive concept of the amount of information in the context of the following example.

Suppose you are planning a trip to Miami, Florida from Minneapolis in the winter time. To determine the weather in Miami, you telephone the Miami weather bureau and receive one of the following forecasts:

1. mild and sunny day,
2. cold day,
3. possible snow flurries.

The amount of information received is obviously different for these messages. The first message contains very little information since the weather in Miami is mild and sunny most of the time. The forecast of a cold day contains more information since it is not an event that occurs often. In comparison, the forecast of snow flurries conveys even more information since the occurrence of snow in Miami is a rare event. *Thus on an* درونی *intuitive basis the amount of information received from the knowledge of occurrence of an event is related to the probability or the* احتمال *likelihood of occurrence of the event.* The message associated with an event least likely to occur contains most information. The above حدس conjecture applies to messages related to any uncertain event, such as the behavior of the stock market. The amount of information in a message depends only on the uncertainty of the underlying event rather than its actual content. We can now formalize this concept in terms of probabilities as follows:

Suppose an information source emits one of q possible messages $m_1, m_2, \ldots, m_q$ with probabilities of occurrence $p_1, p_2, \ldots, p_q$; $p_1 + p_2 + \cdots + p_q = 1$. According to our intuition, the information content or the amount of information in the kth message, denoted by $I(m_k)$, must be inversely related to p_k. Also, to satisfy our intuitive concept of information, $I(m_k)$ must approach 0 as p_k approaches 1. For example, if the forecast in the preceding example said the sun will rise in the east, this does not convey any information since the sun will rise in the east with probability 1. Furthermore, the information content $I(m_k)$ must be nonnegative since each message contains some information. At worst, $I(m_k)$ may be equal to zero. Summarizing these requirements, $I(m_k)$ must satisfy:

$$I(m_k) > I(m_j) \quad \text{if } p_k < p_j \tag{4.1}$$

$$I(m_k) \to 0 \quad \text{as } p_k \to 1 \tag{4.2}$$

تحمیل کردن

$$I(m_k) \geq 0 \quad \text{when } 0 \leq p_k \leq 1 \tag{4.3}$$

Before we start searching for a measure of information that satisfies the above constraints, let us impose one more requirement. Namely, when two *independent* messages are received the total information content is the sum of the information conveyed by each of the two messages. For example, suppose that you read in the newspaper two items of news: (1) scientists have discovered a علاج cure for the common cold and (2) a NASA space کاوشگر probe has found evidence of life on planet Mars. It is reasonable to assume that the two events mentioned in the news items are independent, and that the total information received from the two messages is the same as the sum of the information contained in each of the two news items.

We can apply the same concept to independent messages coming from the same source. For example, the information received in the message, "It will be sunny today and cloudy tomorrow," is the same as the sum of information received in the two messages, "It will be sunny today" and "It will be cloudy tomorrow" (assuming that weather today does not affect weather tomorrow). Mathematically, we can state this requirement by

مستقل

$$\boxed{I(m_k \text{ and } m_j) \overset{\Delta}{=} I(m_k m_j) = I(m_k) + I(m_j)} \tag{4.4}$$

where m_k and m_j are two independent messages.

A continuous function of p_k that satisfies the constraints specified in Equations (4.1)–(4.4) is the logarithmic function and we can define a measure of information as

$$\boxed{I(m_k) = \log(1/p_k)} \tag{4.5}$$

The base for the logarithm in (4.5) determines the unit assigned to the information content. If the natural logarithm base is used, then the unit is *nat*, and if the base is 10, then the unit is *Hartley* or *decit*. When the base is 2, then the unit of information is the familiar *bit*, an اختصار abbreviation for binary digit. Using the binary digit as the unit of information is based on the fact that if two possible binary digits occur with equal probabilities ($p_1 = p_2 = \frac{1}{2}$), then the correct identification of the binary digit conveys an amount of information $I(m_1) = I(m_2) = -\log_2(\frac{1}{2}) = 1$ bit. Unless otherwise specified, we will use the base 2 in our definition of information content.

Example 4.1. A source puts out one of five possible messages during each message interval. The probabilities of these messages are

$$p_1 = \tfrac{1}{2}, \quad p_2 = \tfrac{1}{4}, \quad p_3 = \tfrac{1}{8}, \quad p_4 = \tfrac{1}{16}, \quad p_5 = \tfrac{1}{16}$$

Find the information content of each of these messages. (Observe that the actual meaning or interpretation of the message does not enter into our computation of information content.)

Solution

$$\begin{aligned} I(m_1) &= -\log_2(\tfrac{1}{2}) = 1 \text{ bit} \\ I(m_2) &= -\log_2(\tfrac{1}{4}) = 2 \text{ bits} \\ I(m_3) &= -\log_2(\tfrac{1}{8}) = 3 \text{ bits} \\ I(m_4) &= -\log_2(\tfrac{1}{16}) = 4 \text{ bits} \\ I(m_5) &= -\log_2(\tfrac{1}{16}) = 4 \text{ bits} \end{aligned}$$

4.2.2 Average Information Content (Entropy) of Symbols in Long Independent Sequences

Messages produced by information sources consist of sequences of symbols. While the receiver of a message may interpret the entire message as a single unit, communication systems often have to deal with individual symbols. For example, if we are sending messages in the English language using a teletype, the user at the receiving end is interested mainly in words, phrases, and sentences, whereas the communication system has to deal with individual letters or symbols. Hence, from the point of view of communication systems that have to deal with symbols, we need to define the information content of symbols.

When we attempt to define the information content of symbols, we need to keep the following two factors in mind: First, the instantaneous flow of information in a system may fluctuate widely due to the randomness involved

نوسانی

in the symbol selection. Hence we need to talk about *average information content* of symbols in a long message. Secondly, the statistical dependence of symbols in a message sequence will alter the average information content of symbols. For example, the presence of the letter U following Q in an English word carries less information than the presence of the same letter U following the letter T. We will first define the average information content of symbols assuming the source selects or emits symbols in a statistically independent sequence, with the probabilities of occurrence of various symbols being invariant with respect to time. Later in the chapter we will deal with sources emitting symbols in statistically dependent sequences.

Suppose we have a source that emits one of M possible symbols $s_1, s_2, \ldots, s_M$ in a statistically independent sequence. That is, the probability of occurrence of a particular symbol during a symbol time interval does not depend on the symbols emitted by the source during the preceding symbol intervals. Let $p_1, p_2, \ldots, p_M$ be the probabilities of occurrence of the M symbols, respectively. Now, in a long message containing N symbols, the symbol s_1 will occur on the average p_1N times, the symbol s_2 will occur p_2N times, and in general the symbol s_i will occur p_iN times. Treating individual symbols as messages of length one, we can define the information content of the ith symbol as $\log_2(1/p_i)$ bits. Hence the p_iN occurrences of s_i contributes an information content of $p_iN \log_2(1/p_i)$ bits. The total information content of the message is then the sum of the contribution due to each of the M symbols of the source alphabet and is given by

$$I_{\text{total}} = \sum_{i=1}^{M} Np_i \log_2(1/p_i) \text{ bits}$$

We obtain the *average information per symbol* by dividing the total information content of the message by the number of symbols in the message, as

$$H = \frac{I_{\text{total}}}{N} = \sum_{i=1}^{M} p_i \log_2(1/p_i) \text{ bits/symbol} \tag{4.6}$$

Observe that the definition of H given in Equation (4.6) is based on "time averaging." In order for this definition to be valid for ensemble averages, the source has to be ergodic (see Section 3.5.2).

The expression given in Equation (4.6) was used by Shannon as the starting point in his original presentation of the mathematical theory of communication.

The average information content per symbol is also called the *source entropy* since the expression in (4.6) is similar to the expression for entropy in statistical mechanics. A simple interpretation of the source entropy is the following: On the average, we can expect to get H bits of information per symbol in long messages from the information source even though we cannot say in advance what symbol sequences will occur in these messages.

Example 4.2. Find the entropy of a source that emits one of three symbols A, B, and C in a statistically independent sequence with probabilities $\frac{1}{2}$, $\frac{1}{4}$, and $\frac{1}{4}$, respectively.

Solution. We are given $s_1 = A$, $s_2 = B$, and $s_3 = C$, with $p_1 = \frac{1}{2}$ and $p_2 = p_3 = \frac{1}{4}$. The information content of the symbols are one bit for A, two bits for B, and two bits for C. The average information content per symbol or the source entropy is given by

$$H = \tfrac{1}{2}\log_2(2) + \tfrac{1}{4}\log_2(4) + \tfrac{1}{4}\log_2(4)$$
$$= 1.5 \text{ bits/symbol}$$

(A typical message or symbol sequence from this source could be: *ABCCABAABCABABAACAAB*.)

To explore the dependence of H on the symbol probabilities, let us consider a source emitting two symbols with probabilities p and $1-p$, respectively $(0 < p < 1)$. The entropy for this source is given by

$$H = p \log_2\left(\frac{1}{p}\right) + (1-p)\log_2\left(\frac{1}{1-p}\right) \text{ bits/symbol}$$

It is easy to verify that the maximum value of H is reached when $p = \frac{1}{2}$ $(dH/dp = 0$ requires $\log((1-p)/p) = 0$ and hence $p = \frac{1}{2})$, and $H_{\max}$ is 1 bit/symbol. In general, for a source with an alphabet of M symbols, the *maximum entropy* is attained when the symbol probabilities are equal, that is, when $p_1 = p_2 = \cdots = p_M = 1/M$, and $H_{\max}$ is given by

$$H_{\max} = \log_2 M \text{ bits/symbol} \tag{4.7}$$

It was mentioned earlier that symbols are emitted by the source at a fixed time rate, say r_s symbols/sec. We can bring this time element into the picture and define the average *source information rate* R in bits per second as the product of the average information content per symbol and the symbol rate r_s.

$$R = r_s H \text{ bits/sec} \tag{4.8}$$

اختصار

The abbreviation BPS is often used to denote bits per second.

Example 4.3. A discrete source emits one of five symbols once every millisecond. The symbol probabilities are $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{16}$, respectively. Find the source entropy and information rate.

Solution

$$H = \sum_{i=1}^{5} p_i \log_2\left(\frac{1}{p_i}\right) \text{ bits/symbol}$$
$$= \tfrac{1}{2}\log_2(2) + \tfrac{1}{4}\log_2(4) + \tfrac{1}{8}\log_2(8) + \tfrac{1}{16}\log_2(16) + \tfrac{1}{16}\log_2(16)$$
$$= 0.5 + 0.5 + 0.375 + 0.25 + 0.25$$
$$= 1.875 \text{ bits/symbol}$$

Information rate $R = r_s H$ bits/sec $= (1000)(1.875) = 1875$ bits/sec.

4.2.3 Average Information Content of Symbols in Long Dependent Sequences

The entropy or average information per symbol and the source information rate defined in Equations (4.6) and (4.8) apply to sources that emit symbols in statistically independent sequences. That is, the occurrence of a particular symbol during a symbol interval does not alter the probability of occurrences of symbols during any other symbol intervals. However, nearly all practical sources emit sequences of symbols that are statistically dependent. In telegraphy, for example, the messages to be transmitted consist of a sequence of letters, numerals, and special characters. These sequences, however, are not completely random. In general, they form sentences and have a *statistical structure* of, say, the English language. For example, the letter E occurs more frequently than letter Q; occurrence of letter Q implies that the following letter will most probably be the letter U; the occurrence of a consonant implies that the following letter will most probably be a vowel, and so on. This statistical dependence or structure reduces the amount of information coming out of such a source compared to the amount of information coming from a source emitting the same set of symbols in independent sequences. The problem we address now is one of calculating the information rate for discrete sources that emit dependent sequences of symbols or messages. We will first develop a statistical model for sources emitting symbols in dependent sequences, and then use this model to define the entropy and the information rate for the source.

4.2.4 Markoff Statistical Model for Information Sources

For the purpose of analysis, we will assume that the discrete information source emits a symbol once every T_s seconds. The source puts out symbols belonging to a finite alphabet according to certain probabilities depending, in general, on *preceding symbols* as well as the *particular symbol in question.* A physical system or statistical model of a system that produces such a sequence of symbols governed by a set of probabilities is known as a stochastic or random process. We may consider a discrete source, therefore, to be represented by a random process. Conversely, any random process that produces a discrete sequence of symbols chosen from a finite set may be considered a discrete source. This will include for example, natural written languages such as English and German, and also continuous information sources that have been rendered discrete by sampling and quantization.

We can statistically model the symbol sequences emitted by the discrete source by a class of random processes known as discrete stationary Markoff processes (see Section 3.5.4). The general case can be described as follows:

1. The source is in one of n possible states, $1, 2, \ldots, n$ at the beginning of each symbol interval. The source changes state once during each symbol interval from say i to j. The probability of this transition is p_{ij}, which depends only on the initial state i and the final state j, but does not depend on the states during any of the preceding symbol intervals. The transition probabilities p_{ij} $(i, j = 1, 2, \ldots, n;\ \Sigma_{j=1}^{n} p_{ij} = 1)$ remain constant as the process progresses in time.
2. As the source changes state from i to j it emits a symbol. The particular symbol emitted depends on the initial state i and the transition $i \rightarrow j$.
3. Let $s_1, s_2, \ldots, s_M$ be the symbols of the alphabet, and let $X_1, X_2, \ldots, X_k, \ldots$ be a sequence of random variables where X_k represents the kth symbol in a sequence emitted by the source. Then, the probability that the kth symbol emitted is s_q will depend on the previous symbols $X_1, X_2, \ldots, X_{k-1}$ emitted by the source, that is, s_q is emitted by the source with the conditional probability,

$$P(X_k = s_q | X_1, X_2, \ldots, X_{k-1})$$

4. The residual influence of $X_1, X_2, \ldots, X_{k-1}$ on X_k is represented by the state of the system at the beginning of the kth symbol interval. That is, the probability of occurrence of a particular symbol during the kth symbol interval depends only on the state of the system at the beginning of the symbol interval or

$$P(X_k = s_q | X_1, X_2, \ldots, X_{k-1}) = P(X_k = s_q | S_k) \tag{4.9}$$

where S_k is a discrete random variable representing the state of the system at the beginning of the kth interval. (We use the "states"* to "remember" past history or residual influence in the same context as the use of state variables in systems theory, and states in sequential logic networks.)

5. At the beginning of the first symbol interval, the system is in one of the n possible states $1, 2, \ldots, n$ with probabilities $P_1(1), P_2(1), \ldots, P_n(1)$, respectively $(\Sigma_{i=1}^{n} P_i(1) = 1)$.
6. If the probability that the system is in state j at the beginning of the kth symbol interval is $P_j(k)$, then we can represent the transitions of the system as

$$P_j(k+1) = \sum_{i=1}^{n} P_i(k) p_{ij} \tag{4.10}$$

If we let $P(k)$ be an $n \times 1$ column vector whose ith entry is $P_i(k)$ and let ϕ to be an $n \times n$ matrix whose (i, j)th entry is p_{ij}, then we can rewrite Equation (4.10) in matrix form as

$$P(k+1) = \phi^T P(k)$$

The matrix ϕ is called the probability transition matrix of the Markoff process. The process is called a stationary Markoff process if $P(k) = \phi^T P(k)$ for $k = 1$.

Information sources whose outputs can be modeled by discrete stationary Markoff processes are called discrete stationary Markoff sources.

Discrete stationary Markoff sources are often represented in a graph form where the states are represented by nodes of the graph, and the transition between states is represented by a directed line from the initial to the final state. The transition probabilities and the symbols emitted corresponding to various transitions are usually shown marked along the lines of the graph. An example is shown in Figure 4.1. This example corresponds to a source emitting one of three symbols, A, B, and C. *The probability of occurrence of a symbol depends on the particular symbol in question and the symbol immediately preceding it,* that is, the residual or past influence lasts only for a duration of one symbol. Since the last symbol emitted by the source can be A or B or C, the past history can be represented by three states, one for each of the three symbols of the alphabet. If the system is in state one, then the last symbol emitted by the source was A, and the source now emits letter A with probability $\frac{1}{2}$ and returns to state one, or it emits letter B with probability $\frac{1}{4}$ and goes to state three, or it emits letter C and goes to state two with probability $\frac{1}{4}$. The state transition and symbol generation can also be illustrated using a "tree" diagram. A tree diagram is a planar graph where the nodes correspond to states and branches correspond to transitions. Transition between states occurs once every T_s seconds, where $1/T_s$ is the number of symbols per second emitted by the source. Transition probabilities and symbols emitted

*In general, a discrete stationary source with M letters in the alphabet emitting a symbol sequence with a residual influence lasting q symbols can be represented by n states, where $n \leqslant (M)^q$.

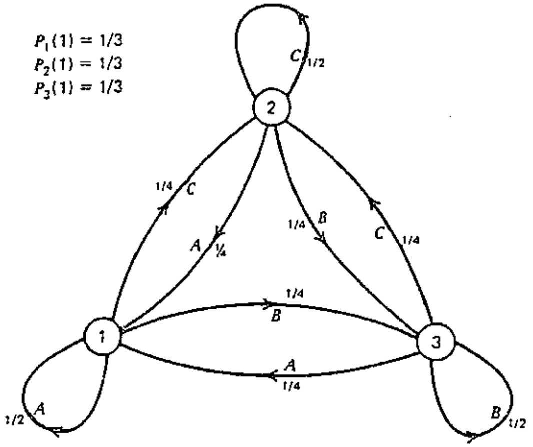


Figure 4.1 Example of a Markoff source.

are shown along the branches. A tree diagram for the source shown in Figure 4.1 is shown in Figure 4.2.

The tree diagram can be used to obtain the probabilities of generating various symbol sequences. For example, the symbol sequence AB can be generated by either one of the following transitions: 1→1→3 or 2→1→3 or 3→1→3. Hence the probability of the source emitting the two-symbol sequence AB is given by

$$\begin{aligned} P(AB) = P(S_1 = 1,\ & S_2 = 1,\ S_3 = 3 \\ \text{or}\quad & S_1 = 2,\ S_2 = 1,\ S_3 = 3) \\ \text{or}\quad & S_1 = 3,\ S_2 = 1,\ S_3 = 3 \end{aligned} \tag{4.11}$$

Since the three transition paths are disjoint, we get

$$\begin{aligned} P(AB) = {} & P(S_1 = 1,\ S_2 = 1,\ S_3 = 3) \\ & + P(S_1 = 2,\ S_2 = 1,\ S_3 = 3) \\ & + P(S_1 = 3,\ S_2 = 1,\ S_3 = 3) \end{aligned} \tag{4.12}$$

Using the chain rule of probability we can rewrite the first term on the right-hand side of Equation (4.12) as

$$\begin{aligned} & P\{S_1 = 1,\ S_2 = 1,\ S_3 = 3\} \\ & \quad = P(S_1 = 1)P(S_2 = 1 | S_1 = 1)P(S_3 = 3 | S_1 = 1,\ S_2 = 1) \\ & \quad = P(S_1 = 1)P(S_2 = 1 | S_1 = 1)P(S_3 = 3 | S_2 = 1) \end{aligned} \tag{4.13}$$

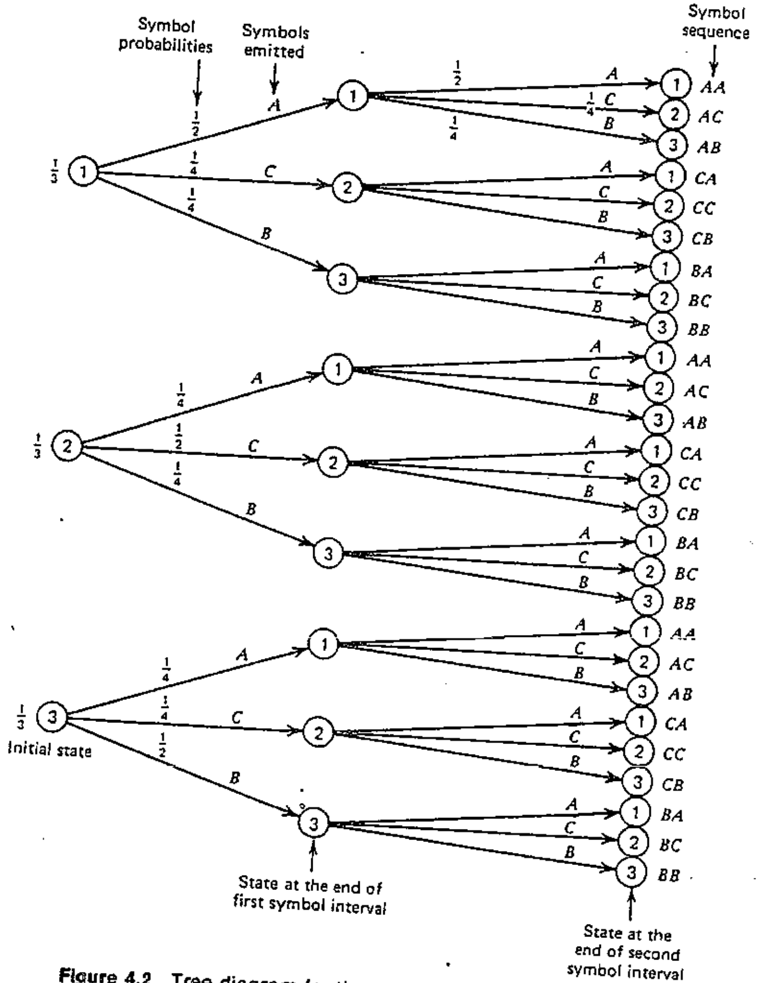


Figure 4.2 Tree diagram for the source shown in Figure 4.1.

The last step is based on the fact that the transition probability to S_3 depends on S_2 but not on how the system got to S_2 (i.e., the Markoff property).

The right-hand side of the previous equation is the product of probabilities shown along the branches representing the transition path and the probability of being at state 1 at the starting point. Other terms on the right-hand side of

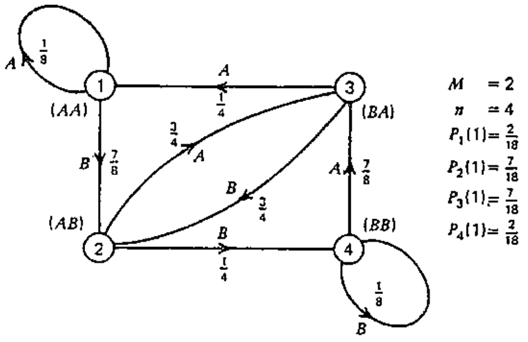


Figure 4.3 Another example of a Markoff source.

Equation (4.12) can be similarly evaluated, and $P(AB)$ is given by

$$\begin{aligned} P(AB) &= (\tfrac{1}{3})(\tfrac{1}{2})(\tfrac{1}{4}) + (\tfrac{1}{3})(\tfrac{1}{4})(\tfrac{1}{4}) + (\tfrac{1}{3})(\tfrac{1}{4})(\tfrac{1}{4}) \\ &= \tfrac{1}{12} \end{aligned}$$

Similarly, the probabilities of occurrence of other symbol sequences can be computed. In general, the probability of the source emitting a particular symbol sequence can be computed by summing the product of probabilities in the tree diagram along all the paths that yield the particular sequence of interest.

The model shown in Figure 4.1 corresponds to a source in which the residual influence lasts over one symbol interval. Shown in Figure 4.3 is a source where the probability of occurrence of a symbol depends not only on the particular symbol in question, but also on the two symbols preceding it. It is easy to verify that if the system is in state two, at the beginning of a symbol interval, then the last two symbols emitted by the source were AB, and so on.

It is left as an exercise for the reader to draw the tree diagram for this source and to calculate the probabilities of some typical symbol sequences (Problem 4.11).

4.2.5 Entropy and Information Rate of Markoff Sources

In this section we will define the entropy and information rate of Markoff sources. We will assume that the source can be modeled as a discrete finite-state Markoff process. Furthermore, we will assume the process to be *ergodic* (Chapter 3, Section 3.4) so that time averages can be applied. The ergodic assumption implies that the process is stationary, and hence $P_i(k) = P_i(k+j)$ for any values of k and j. In other words, the probability of being in state i at the beginning of the first symbol interval is the same as the probability of being in state i at the beginning of the second symbol interval, and so on. The probability of going from state i to j also does not depend on time.

We define the entropy of the source as a weighted average of the entropy of the symbols emitted from each state, where the entropy of state i, denoted by H_i, is defined as the average information content of the symbols emitted from the i-th state:

$$H_i = \sum_{j=1}^{n} p_{ij} \log_2\left(\frac{1}{p_{ij}}\right) \text{ bits/symbol} \tag{4.14}$$

The entropy of the source is then the average of the entropy of each state. That is,

$$\begin{aligned} H &= \sum_{i=1}^{n} P_i H_i \\ &= \sum_{i=1}^{n} P_i \left[\sum_{j=1}^{n} p_{ij} \log_2\left(\frac{1}{p_{ij}}\right)\right] \text{ bits/symbol} \end{aligned} \tag{4.15}$$

where P_i is the probability that the source is in state i. The average information rate R for the source is defined as

$$R = r_s H \text{ bits/sec} \tag{4.16}$$

where r_s is the number of state transitions per second or the symbol rate of the source.

The entropy H defined in Equation (4.15) carries the same meaning as H defined in Equation (4.6). In both cases, we can expect the source output to convey H bits of information per symbol in long messages. This is precisely stated in the following theorem.

Theorem 4.1

Let $P(m_i)$ be the probability of a sequence m_i of N symbols from the source. Let

$$G_N = -\frac{1}{N} \sum_i P(m_i) \log_2 P(m_i) \tag{4.17}$$

where the sum is over all sequences m_i containing N symbols. Then G_N is monotonic decreasing function of N and

یکنواخت

$$\lim_{N \to \infty} G_N = H \text{ bits/symbol} \tag{4.18}$$

A detailed proof of this theorem can be found in Reference 1. We w illustrate the concept stated in Theorem 4.1 by the following example.

Example 4.4. Consider an information source modeled by a discrete ergodic Markoff random process whose graph is shown in Figure 4.4. Find the source entropy H and the average information content per symbol in messages containing one, two, and three symbols, that is, find G_1, G_2, and G_3.

Solution. The source shown above emits one of three symbols A, B, and C. The tree diagram for the output of the source is shown in Figure 4.5 and Table 4.1 lists the various symbol sequences and their probabilities. To illustrate how these messages and their probabilities are generated, let us consider the sequence CCC. There are two paths on the graph that terminate in CCC corresponding to the transition sequences $1 \to 2 \to 1 \to 2$ and $2 \to 1 \to 2 \to 1$. The probability of the path $1 \to 2 \to 1 \to 2$ is given by the product of the probability that the system is in state one initially, and the probabilities of the transitions $1 \to 2$, $2 \to 1$, and $1 \to 2$. These probabilities are $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{4}$, and $\frac{1}{4}$, respectively, and hence the path probability is 1/128. Similarly, the probability of the second path can be calculated as 1/128. Hence, the probability of the sequence CCC is given by the sum of the two paths as 2/128.

From the definition of H_i (Equation (4.14)) we get

$$H_1 = \tfrac{1}{4}\log_2(4) + \tfrac{3}{4}\log_2(\tfrac{4}{3}) = 0.8113$$
$$H_2 = \tfrac{1}{4}\log_2(4) + \tfrac{3}{4}\log_2(\tfrac{4}{3}) = 0.8113$$

and using Equation (4.15) we obtain the source entropy as

$$H = (\tfrac{1}{2})(0.8113) + (\tfrac{1}{2})(0.8113) = 0.8113 \text{ bits/symbol}$$

Let us now calculate the average information content per symbol in messages containing two symbols. There are seven messages. The information contents of these messages are $I(AA) = I(BB) = 1.83$, $I(BC) = I(AC) = I(CB) =$

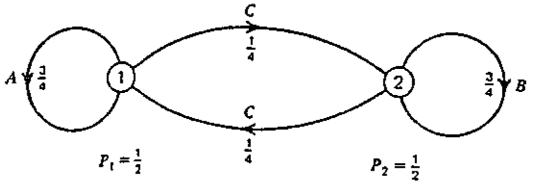


Figure 4.4 Source model for Example 4.4.

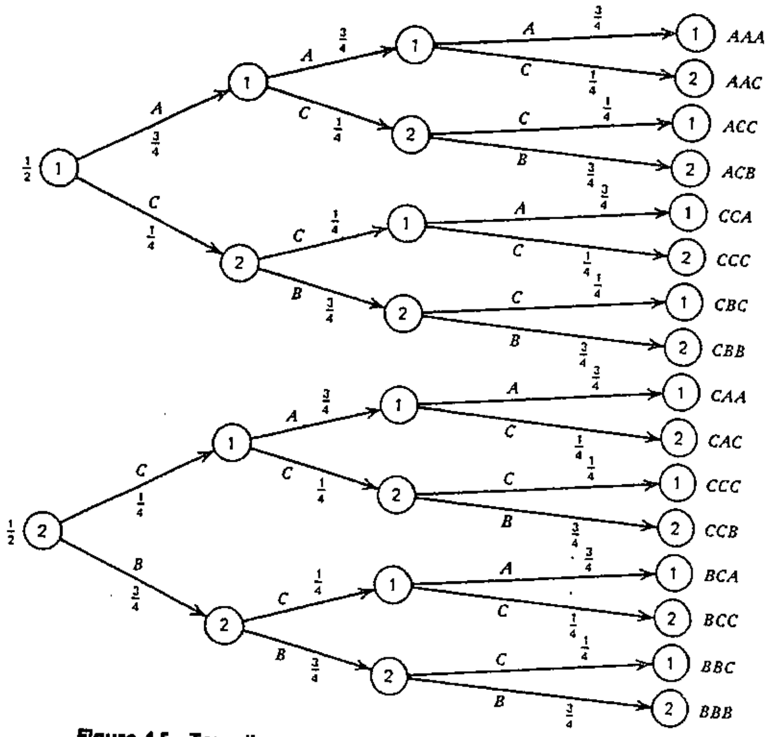


Figure 4.5 Tree diagram for the source shown in Figure 4.4.

Table 4.1. Messages of length 1, 2, and 3 and their probabilities

Messages of length 1	Messages of length 2	Messages of length 3
A (3/8)	AA (9/32)	AAA (27/128)
B (3/8)	AC (3/32)	AAC (9/128)
C (1/4)	CB (3/32)	ACC (3/128)
	CC (2/32)	ACB (9/128)
	BB (9/32)	BBB (27/128)
	BC (3/32)	BBC (9/128)
	CA (3/32)	BCC (3/128)
		BCA (9/128)
		CCA (3/128)
		CCB (3/128)
		CCC (2/128)
		CBC (3/128)
		CAC (3/128)
		CBB (9/128)
		CAA (9/128)

$I(CA)=3.4150$, and $I(CC)=4.0$ bits. The average information content of these messages is given by the sum of the products of the information content of each message and its respective probability. This can be computed as 2.5598 bits. Now we can obtain the average information content per symbol in messages containing two symbols by dividing the average information content of the messages by the number of symbols in these messages, that is,

$$G_2 = 2.5598/2 = 1.2799 \text{ bits/symbol}$$

In a similar fashion, we can obtain the values of G_1 and G_3. The reader can easily verify that

$$G_1 = 1.5612 \text{ bits/symbol}$$

$$G_3 = 1.0970 \text{ bits/symbol}$$

Thus,

$$G_1 \geqslant G_2 \geqslant G_3 \geqslant H$$

as stated in Theorem 4.1.

The preceding example illustrates the basic concept that the average information content per symbol from a source emitting symbols in a dependent sequence decreases as the message length increases. Alternatively, the average number of bits per symbol needed to represent a message decreases as the message length increases. The decrease in entropy is due to the structure of the messages—messages that are highly structured usually convey less information per symbol than messages containing the same number of symbols when the symbols are chosen independently.

In the next section we will discuss a source coding technique that takes advantage of the statistical structure of the source to reduce the average number of bits per symbol needed to represent the output of an information source. But, before, we discuss source coding techniques, let us take a brief look at how the statistical model of a source is constructed.

The development of a model for an information source consists of two parts: (1) the development of the model structure and (2) estimation of the parameters of the model. The structure of the model is usually derived from knowledge about the physical nature of the source. Parameter values are obtained through the use of statistical estimation procedures. In some cases, both the structure and the parameters of the source model can be derived from test data using estimation techniques.

The test data used for estimating the parameters of a source model can be derived either from the simultaneous recordings of the outputs of a number of identical sources for a short time or from the recording of the output of a single source for a long time period. Estimates based on data from a large number of sources are called ensemble estimates, while estimates based on the output of a single source are called time-averaged estimates. These two estimates will be equal if the source is ergodic. An excellent treatment of how to collect and analyze random data and to estimate parameter values can be found in Reference 2. We will illustrate through an example the principle involved in such procedures.

Example 4.5. We want to design a system to report the heading of a collection of 400 cars. The heading is to be quantized into three levels: heading straight (S), turning left (L), and turning right (R). This information is to be transmitted every second. Based on the test data given below, construct a model for the source and calculate the source entropy.

1. On the average, during a given reporting interval, 200 cars were heading straight, 100 were turning left, and 100 cars were turning right.
2. Out of 200 cars that reported heading straight during a reporting period, 100 of them (on the average) reported going straight during the next reporting period, 50 of them reported turning left during the next period, and 50 of them reported turning right during the next period.
3. On the average, out of 100 cars that reported as turning during a signaling period, 50 of them continued their turn during the next period and the remaining headed straight during the next reporting period.
4. The dynamics of the cars did not allow them to change their heading from left to right or right to left during subsequent reporting periods.

Solution. The source model for this process can be developed as follows:

1. Since the past history or residual influence lasts one reporting interval, we need only three states to "remember" the last symbol emitted by the source. The state probabilities are given as (statement one)

$$P_1 = \tfrac{1}{4}, \quad P_2 = \tfrac{1}{2}, \quad P_3 = \tfrac{1}{4}$$

2. The transition probabilities are

$$P\left(\begin{matrix}\text{present}\\ \text{heading}\end{matrix} = S \,\middle|\, \begin{matrix}\text{previous}\\ \text{heading}\end{matrix} = S\right) = P(S|S) = 0.5$$

$$P(S|S) = P(R|R) = P(L|L) = 0.5$$

$$P(L|S) = P(R|S) = 0.25$$

$$P(S|L) = P(S|R) = 0.5$$

$$P(L|R) = P(R|L) = 0$$

3. The state diagram is shown in Figure 4.6.

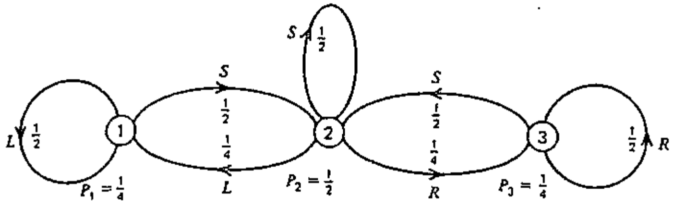


Figure 4.6 Source model for Example 4.5.

The data discussed in the preceding example could have been obtained by monitoring the heading of a single car over an extended period of time instead of monitoring 400 cars. We can obtain good estimates of statistical parameters from a single time history record if the process is ergodic. Here, we will estimate the probability of a car heading straight by the portion of time the car was heading straight during a long monitoring period (as opposed to the ratio of 400 cars heading straight at a given time in the previous method).

The information rate for this source can be computed as follows:

H_1 = entropy of state one
$= -\sum p_{1j} \log_2(p_{1j})$, where the sum is over all symbols emitted from state one
$= \frac{1}{2}\log_2 2 + \frac{1}{2}\log_2 2 = 1$ bit/symbol
$H_2 = \frac{1}{2}\log_2 2 + \frac{1}{4}\log_2 4 + \frac{1}{4}\log_2 4$
$= 1.5$ bits/symbol
$H_3 = 1$ bit/symbol, and hence
$H = P_1 H_1 + P_2 H_2 + P_3 H_3$
$= \frac{1}{4}(1) + \frac{1}{2}(1.5) + \frac{1}{4}(1)$
$= 1.25$ bits/symbol, and
$R = 1.25$ bits/sec (per car)

The values of probabilities obtained from data are estimated values and are not the actual values. The sample size and the type of estimator used will determine the bias (if any) and the variance of the estimator for each of the parameters.

4.3 ENCODING OF THE SOURCE OUTPUT

Source encoding is the process by which the output of an information source is converted into a binary sequence. The functional block that performs this task in a communication system is called the *source encoder*. The input to the source encoder is the symbol sequence emitted by the information source. The source encoder assigns variable length binary code words to blocks of symbols and produces a binary sequence as its output. If the encoder operates on blocks of N symbols in an optimum way, it will produce an average bit rate of G_N bits/symbol, where G_N is defined in Theorem 4.1. In general, the average bit rate of the encoder will be greater than G_N due to practical constraints. As the block length N is increased, the average output bit rate per symbol will decrease and in the limiting case as $N \to \infty$, the bit rate per symbol will approach G_N and G_N will approach H. Thus, with a large block size the output of the information source can be encoded into a binary sequence with an average bit rate approaching R, the source information rate.

The performance of the encoder is usually measured in terms of the *coding efficiency* that is defined as the ratio of the source information rate and the average output bit rate of the encoder. There are many algorithms available for designing a source encoder. The following section deals with a simple, yet most powerful, source coding algorithm given by Shannon.

4.3.1 Shannon's Encoding Algorithm

The design of the source encoder can be formulated as follows:

> The input to the source encoder consists of one of q possible messages, each message containing N symbols. Let $p_1, p_2, \ldots, p_q$ be the probabilities of these q messages. We would like to code (or replace) the ith message m_i using a unique binary code word c_i of length n_i bits, where n_i is an integer. Our objective is to find n_i and c_i for $i = 1, 2, \ldots, q$ such that the average number of bits per symbol $\hat{H}_N$ used in the coding scheme is as close to G_N as possible. In other words, we want
>
> $$\hat{H}_N = \frac{1}{N}\sum_{i=1}^{q} n_i p_i \to \frac{1}{N}\sum_{i=1}^{q} p_i \log_2\left(\frac{1}{p_i}\right)$$

Several solutions have been proposed to the above problem, and the algorithm given by Shannon (and Fano) is stated below.

> Suppose the q messages $m_1, m_2, \ldots, m_q$ are arranged in order of decreasing probability such that $p_1 \geq p_2 \geq \cdots \geq p_q$. Let $F_i = \sum_{k=1}^{i-1} p_k$, with $F_1 = 0$. Let n_i be an integer such that
>
> $$\log_2(1/p_i) \leq n_i < 1 + \log_2(1/p_i) \qquad (4.19)$$

Then, the code word for the message m_i is the binary expansion* of the fraction F_i up to n_i bits, that is,

$$c_i = (F_i)_{\text{binary } n_i \text{ bits}}$$

This algorithm yields a source encoding procedure that has the following properties:

1. Messages of high probability are represented by short code words and those of low probability are represented by long code words. This can be easily verified using the inequality stated in (4.19).
2. The code word for m_i will differ from all succeeding code words in one or more places and hence it is possible to decode messages uniquely from their code words. We can prove this by rewriting inequality (4.19) as

$$\frac{1}{2^{n_i}} \leq p_i < \frac{1}{2^{n_i-1}}$$

Hence the binary expansion of F_i will differ from all succeeding ones in one or more places. For example, F_i and F_{i+1} will differ in the n_ith bit since $p_i \geq 1/2^{n_i}$. Hence the code word for m_{i+1} will differ from m_i in at least one bit position or more.

3. The average number of bits per symbol used by the encoder is bounded by

$$G_N \leq \hat{H}_N < G_N + 1/N \tag{4.20}$$

This bound can be easily verified as follows:

From (4.19) we have

$$\log_2(1/p_i) \leq n_i < 1 + \log_2(1/p_i)$$

Multiplying throughout by p_i and summing over i, we obtain

$$\sum_{i=1}^{q} p_i \log_2(1/p_i) \leq \sum_{i=1}^{q} n_i p_i < 1 + \sum_{i=1}^{q} p_i \log_2(1/p_i)$$

or

$$\frac{1}{N}\sum_{i=1}^{q} p_i \log_2\left(\frac{1}{p_i}\right) \leq \frac{1}{N}\sum_{i=1}^{q} p_i n_i < \frac{1}{N} + \frac{1}{N}\sum_{i=1}^{q} \log_2\left(\frac{1}{p_i}\right)$$

or

$$G_N \leq \hat{H}_N < 1/N + G_N$$

Hence as $N \to \infty$, $G_N \to H$ and $\hat{H}_N \to H$.

*Binary fraction $\quad .b_1b_2b_3\ldots b_k = \frac{b_1}{2} + \frac{b_2}{2^2} + \frac{b_3}{2^3} + \cdots + \frac{b_k}{2^k}$

where $b_i = 0$ or 1.

The *rate efficiency* e of the encoder using blocks of N symbols is defined as

$$e = H/\hat{H}_N \tag{4.21}$$

The following example illustrates the concepts involved in the design of a source encoder.

Example 4.6. Design a source encoder for the information source given in Example 4.4. Compare the average output bit rate and efficiency of the coder for $N = 1$, 2, and 3.

Solution. Let us first design the encoder with a block size $N = 3$. From Example 4.4 we know that the source emits 15 distinct three-symbol messages. These messages and their probabilities are shown in columns 1 and 2 of Table 4.2; the messages are arranged in column 1 according to decreasing order of probabilities. The number of bits n_1 assigned to message m_1 is bounded by

$$\log_2\left(\frac{128}{27}\right) \leq n_1 < 1 + \log_2\left(\frac{128}{27}\right)$$

or

$$2.245 \leq n_1 < 3.245$$

Table 4.2. Source encoder design for Example 4.6

Messages m_i	p_i	n_i	F_i	Binary expansion of F_i	Code word c_i
AAA	27/128	3	0	.0000000	000
BBB	27/128	3	27/128	.0011011	001
CAA	9/128	4	54/128	.0110110	0110
CBB	9/128	4	63/128	.0111111	0111
BCA	9/128	4	72/128	.1001000	1001
BBC	9/128	4	81/128	.1010001	1010
AAC	9/128	4	90/128	.1011010	1011
ACB	9/128	4	99/128	.1100011	1100
CBC	3/128	6	108/128	.1101100	110110
CAC	3/128	6	111/128	.1101111	110111
CCB	3/128	6	114/128	.1110010	111001
CCA	3/128	6	117/128	.1110101	111010
BCC	3/128	6	120/128	.1111000	111100
ACC	3/128	6	123/128	.1111011	111101
CCC	2/128	6	126/128	.1111110	111111

$\Sigma\, p_i n_i = 3.89 \quad \hat{H}_3 = \frac{3.89}{3} \approx 1.30$ bits/symbol

Since n_1 has to be an integer, the above inequality yields $n_1 = 3$ bits. The code word c_1 is generated from $F_1 \triangleq 0$. Hence, $c_1 = 000$. For m_2, it is easy to verify that $n_2 = 3$ bits and $F_2 = \Sigma_{i=1}^{1} p_i$ or $F_2 = 27/128$. The binary expansion of 27/128 is .0011011. Truncating this expansion to 3 bits, we obtain the code word 001 for m_2. The complete design of the encoder for $N = 3$ is shown in Table 4.2.

It can be easily verified that the average number of bits per symbol used by the encoder is 1.30 bits/symbol. Table 4.3 summarizes the characteristics of the encoder for $N = 1$ and 2.

The performance of the encoder is summarized in Table 4.4. The numbers in Table 4.4 show that the average output bit rate of the encoder decreases as N increases, and hence the efficiency of the encoder increases as N increases. Also we can verify that

$$\hat{H}_N < G_N + 1/N \quad \text{for } N = 1, 2, 3$$

To illustrate how the encoder operates, let us consider a symbol string $ACBBCAAACBBB$ at the encoder input. If the encoder uses a block size $N = 3$, then the output of the encoder can be obtained by replacing successive groups of three input symbols by the code words shown in Table 4.2 as

$$\begin{matrix} ACB & BCA & AAC & BBB \\ 1100 & 1001 & 1011 & 001 \end{matrix}$$

The same symbol string will be encoded as 100101110100101001 if the encoder operates on two symbols at a time with code words listed in Table 4.3. The decoding is accomplished by starting at the left-most bit and matching groups of bits with the code words listed in the code table.

For the $N = 3$ example, we take the first 3-bit group 110 (this is the shortest code word) and check for a matching word in Table 4.2. Finding none, we try the first 4-bit group 1100, find a match, and decode this group as ACB. Then the procedure is repeated beginning with the fifth bit to decode the rest of the symbol groups. The reader can verify that the decoding can be done easily by knowing the code word lengths a priori if no errors occur in the bit string in the transmission process.

Bit errors in transmission lead to serious decoding problems. For example, if the bit string 11001001101111001 ($N = 3$) was received at the decoder input with one bit error as 1101100110111001, the message will be decoded as $CBCCAACCB$ instead of $ACBBCAAACBCA$. This type of error is a major disadvantage of an encoder using variable length code words. Another disadvantage lies in the fact that output data rates measured over short time periods will fluctuate widely. To avoid this problem, buffers of large length will be needed at both the encoder and the decoder to store the variable rate bit stream if a fixed output rate is to be maintained.

Some of the above difficulties can be resolved by using *fixed length* code words at the expense of a slight increase in data rate. For example, with $N = 3$ we can encode the output of the source discussed in the preceding example using 4-bit code words $0000, 0001, \ldots, 1110$. The output data rate now is *fixed* at 1.333 bits/symbol compared to an *average output* data rate of 1.30 bits/symbol for the variable length scheme discussed before. The encoder/decoder structure using fixed length code words will be very simple

Table 4.3. Encoder for $N = 1$ and $N = 2$

Message	p_i	n_i	C_i
$N = 1$			
A	3/8	2	00
B	3/8	2	01
C	1/4	2	11
$\hat{H}_1 = 2$ bits/symbol			
$N = 2$			
AA	9/32	2	00
BB	9/32	2	01
AC	3/32	4	1001
CB	3/32	4	1010
BC	3/32	4	1100
CA	3/32	4	1101
CC	2/32	4	1111
$\hat{H}_2 = 1.44$ bits/symbol			

Table 4.4. Performance of the encoder for Example 4.6

Average number of bits per symbol used	$N = 1$	$N = 2$	$N = 3$
$\hat{H}_N$	2	1.44	1.30
G_N	1.561	1.279	1.097
$G_N + 1/N$	2.561	1.779	1.430
Efficiency $= \dfrac{H}{\hat{H}_N}$	40.56%	56.34%	62.40%
$H = .8113$			

compared to the complexity of an encoder/decoder using variable length code words. Also single bit errors lead to single block errors when fixed length code words are used. These two advantages more than make up for the slight increase in data rate from 1.30 bits/symbol to 1.33 bits/symbol.

Another important parameter in the design of encoders is the delay involved in decoding a symbol. With large block sizes, the first symbol in the block cannot be decoded until the bit string for the entire block is received by the decoder. The average delay will be $N/2$ symbols for a block size of N symbols. The time delay $(N/2)T_s$ seconds, where $1/T_s$ is the number of symbols emitted by a source, may be unacceptable in some real time applications.

It must be pointed out here that the encoding algorithm presented in the preceding pages is only one of many encoding algorithms for representing a source output. Other encoding procedures such as the one given by Huffman yield the shortest average word length. These schemes are more difficult to implement and the interested reader is referred to Abramson's book on information theory [2] for details.

It is also possible to represent the output of an information source using code words selected from an alphabet containing more than two letters. The design of source encoder using nonbinary code words is rather involved. The interested reader is referred to the books of Abramson, Wozencraft and Jacobs, and Gallager [2–4].

Having developed the concept of information rate for sources, we now turn our attention to the problem of evaluating the maximum rate at which reliable information transmission can take place over a noisy channel.

4.4 COMMUNICATION CHANNELS

We can divide a practical communication system into a transmitter, physical channel, or transmission medium, and a receiver. The transmitter consists of an encoder and a modulator, while the receiver consists of a demodulator and a decoder. The term "communication channel" carries different meanings and characterizations depending on its terminal points and functionality. Between points c and g in the system shown in Figure 4.7 we have a discrete channel, often referred to as a *coding channel*, that accepts a sequence of symbols at its input and produces a sequence of symbols at its output. This channel is completely characterized by a set of transition probabilities p_{ij}, where p_{ij} is the probability that the channel output is the jth symbol of the alphabet when the channel input is the ith symbol. These probabilities will depend on the parameters of the modulator, transmission media, noise, and demodulator However, this dependence is transparent to a system designer who is concerned with the design of the digital encoder and decoder.

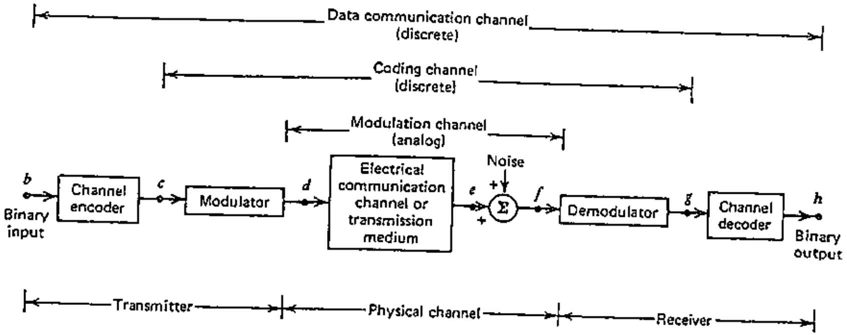


Figure 4.7 Characterization of a binary communication channel.

The communication channel between points d and f in the system provides the electrical connection between the transmitter and the receiver. The input and output are analog electrical waveforms. This portion of the channel is often called a continuous or *modulation channel*. Examples of analog electrical communication channels are voiceband and wideband telephone systems, high frequency radio systems, and troposcatter systems. These channels are subject to several varieties of impairments. Some are due to amplitude and frequency response variations of the channel within its passband. Other channel impairments are due to variations of channel characteristics with time and nonlinearities in the channel. All of these result in the channel modifying the input signal in a deterministic (although not necessarily a known) fashion. In addition, the channel can also corrupt the signal statistically due to various types of additive and multiplicative noise and *fades* (random attenuation changes within the transmission medium). All of these impairments introduce errors in data transmission and limit the maximum rate at which data can be transferred over the channel.

In the following sections we will develop simple mathematical models for discrete communication channels and develop the concept of capacity of a discrete communication channel. The channel capacity is one of the most important parameters of a data communication system since it represents the maximum rate at which data can be transferred between two points in the system, with an arbitrarily small probability of error. After we deal with discrete channels, we will discuss the Shannon–Hartley theorem, which defines the capacity of certain types of continuous channels.

4.5 DISCRETE COMMUNICATION CHANNELS

The communication channel between points c and g in Figure 4.7 is discrete in nature. In the general case, the input to the channel is a symbol belonging to an alphabet of M symbols. The output of the channel is a symbol belonging to the same alphabet of M input symbols. Due to errors in the channel, the output symbol may be different from the input symbol during some symbol intervals. Errors are mainly due to the noise in the analog portion of the communication channel. The discrete channel is completely modeled by a set of probabilities p_i^t $(i = 1, 2, \dots, M)$ and p_{ij} $(i, j = 1, 2, \dots, M)$. p_i^t is the probability that the input to the channel is the ith symbol of the alphabet and p_{ij} is the probability that the ith symbol is received as the jth symbol of the alphabet at the output of the channel. Channels designed to transmit and receive one of M possible symbols are called discrete M-ary channels $(M > 2)$. In the binary case we can statistically model the digital channel as shown in Figure 4.8.

The input to the channel is a binary valued discrete random variable X, and the two nodes on the left-hand side of the graph in Figure 4.8 represent the values 0 and 1 of the random variable X. The output of the channel is also a binary valued random variable Y and its values are shown marked at the nodes on the right-hand side of the graph. Four paths connect the input nodes to the output nodes. The path on the top portion of the graph represents an input 0 and a correct output 0. The diagonal path from 0 to 1 represents an input bit 0 appearing incorrectly as 1 at the output of the channel due to noise. Errors occur in a random fashion and we can statistically model the occurrence of errors by assigning probabilities to the paths shown in Figure 4.8. To simplify the analysis, we will assume that the occurrence of an error during a bit interval does not affect the behavior of the system during other bit intervals (i.e., we will assume the channel to be *memoryless*).

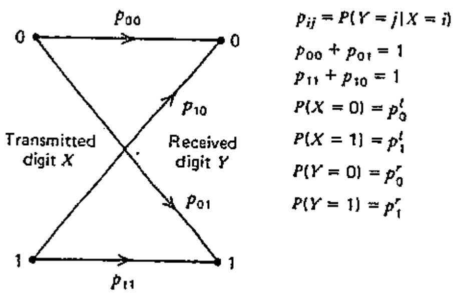


Figure 4.8 Model of a discrete channel.

Letting $P(X = 0) = p_0^t$, $P(X = 1) = p_1^t$, $P(Y = 0) = p_0^r$, $P(Y = 1) = p_1^r$, we have the following relationships:

$$\begin{aligned} P(\text{error}) = P_e = P(X \neq Y) &= P(X = 0, Y = 1) + P(X = 1, Y = 0) \\ &= P(Y = 1 \mid X = 0)P(X = 0) + P(Y = 0 \mid X = 1)P(X = 1) \end{aligned}$$

or

$$P_e = p_0^t p_{01} + p_1^t p_{10} \tag{4.22}$$

Also, p_0^r and p_1^r can be expressed as

$$\begin{aligned} p_0^r &= p_0^t p_{00} + p_1^t p_{10} \\ p_1^r &= p_0^t p_{01} + p_1^t p_{11} \end{aligned} \tag{4.23}$$

The channel is called a *binary symmetric channel* (BSC) if $p_{00} = p_{11} = p$. The only parameter needed to characterize a BSC is p.

We can extend our model to the general case where the channel input X can assume M values $(M > 2)$. There are commercial modems available today where up to eight distinct levels or waveforms are transmitted over the channel. Figure 4.9 shows a model for the general case. Analysis of this channel is similar to the analysis of the binary channel discussed before. For example,

$$p_j^r = \sum_{i=1}^{M} p_i^t p_{ij}$$

and

$$P(\text{error}) = P_e = \sum_{i=1}^{M} p_i^t \left[\sum_{\substack{j=1 \\ j \neq i}}^{M} p_{ij} \right] \tag{4.24}$$

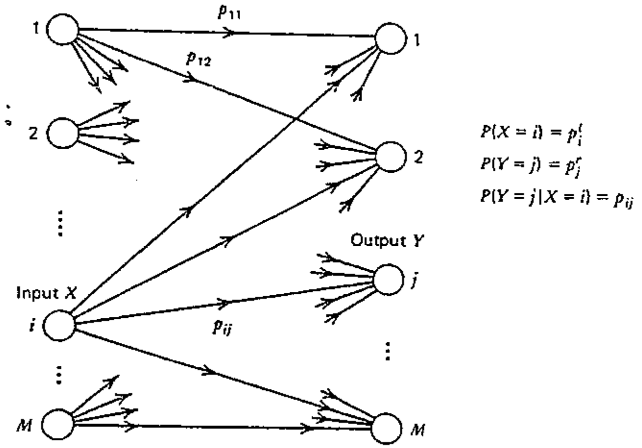


Figure 4.9 Model of an M-ary discrete memoryless channel.

In a discrete memoryless channel such as the one shown in Figure 4.9, there are two statistical processes at work: the input to the channel and the noise. Thus there are a number of entropies or information contents that we can calculate. First, we have the entropy of the input X defined as

$$H(X) = -\sum_{i=1}^{M} p_i^t \log_2(p_i^t) \text{ bits/symbol} \tag{4.25}$$

where p_i^t is the probability that the ith symbol of the alphabet is transmitted. Similarly, we can also define the entropy of the output Y as

$$H(Y) = -\sum_{i=1}^{M} p_i^r \log_2(p_i^r) \text{ bits/symbol} \tag{4.26}$$

where p_i^r denotes the probability that the output of the channel is the ith symbol of the alphabet. $H(Y)$ represents the average number of bits per symbol needed to encode the output of the channel. We can also define a conditional entropy $H(X|Y)$, called *equivocation*, as

$$H(X|Y) = -\sum_{i=1}^{M}\sum_{j=1}^{M} P(X=i, Y=j)\log_2(P(X=i|Y=j)) \tag{4.27}$$

and a joint entropy $H(X, Y)$ as

$$H(X, Y) = -\sum_{i=1}^{M}\sum_{j=1}^{M} P(X=i, Y=j)\log_2 P(X=i, Y=j) \tag{4.28}$$

The conditional entropy $H(X|Y)$ represents how uncertain we are of X, on the average, when we know Y. The reader can verify the following relationships between $H(X)$, $H(Y)$, $H(X|Y)$, $H(Y|X)$, and $H(X, Y)$:

$$\begin{aligned} H(X, Y) &= H(X|Y) + H(Y) \\ &= H(Y|X) + H(X) \end{aligned} \tag{4.29}$$

where

$$H(Y|X) = -\sum_{i=1}^{M}\sum_{j=1}^{M} P(X=i, Y=j)\log_2(P(Y=j|X=i))$$

For a BSC, $P(X=i|Y=i)$ $(i=0, 1)$ measures the uncertainty about the transmitted bit based on the received bit. The uncertainty is minimum when $P(X=i|Y=i)=1$ for $i=0, 1$, that is, an errorless channel. The uncertainty is maximum when $P(X=i|Y=i)=\frac{1}{2}$ for $i=0, 1$. If we define the uncertainty as $-\log_2[P(X=i|Y=i)]$, then we have one bit of uncertainty when the output is independent of the input. When we have one bit of uncertainty associated with each received bit, the received value of the bit does not convey any information!

The conditional entropy $H(X|Y)$ is an *average measure of uncertainty* about X when we know Y. In one extreme we can have Y and X related in a one-to-one manner such as $Y=X$. For this case, there is no uncertainty about X when we know Y; $P(X=i|Y=j)=\delta_{ij}$, where δ_{ij} is the delta function that is 0 for $i \neq j$ and 1 for $i=j$. We can easily verify that $H(X|Y)=0$ when $Y=X$. In the context of a communication channel $Y=X$ represents an errorless channel, and there is no uncertainty about the input when the output is known. Alternatively, we can say that no information is lost in the channel since the output is uniquely related to the input. As another example, let us consider a communication channel that is so noisy that the output is statistically independent of the input. In this case we can easily verify that $H(X, Y) = H(X) + H(Y)$, and $H(X|Y) = H(X)$, that is, Y does not contain any information about X (see Problem 4.19).

4.5.1 Rate of Information Transmission Over a Discrete Channel

In the case of an M-ary discrete memoryless channel accepting symbols at the rate of r_s symbols/sec, the average amount of information per symbol going into the channel is given by the entropy of the input random variable X as

$$H(X) = -\sum_{i=1}^{M} p_i^t \log_2 p_i^t \tag{4.30}$$

In Equation (4.30) we have assumed that the symbols in the sequence at the input to the channel occur in a statistically independent fashion. The average rate at which information is going into the channel is given by

$$D_{in} = H(X) r_s \text{ bits/sec} \tag{4.31}$$

Due to errors, it is not in general possible to reconstruct the input symbol sequence with certainty by operating on the received sequence. Hence we can say that some information is lost due to errors. Before we attempt to define the amount of information that is "lost", let us consider the following example.

Suppose there are two possible symbols 0 and 1 that are transmitted at a rate of 1000 symbols or bits per second with $p_0^t = \frac{1}{2}$ and $p_1^t = \frac{1}{2}$. The source information rate and D_{in} at the input to the channel are 1000 bits/sec. Let us assume that the channel is symmetric with the probability of errorless transmission p equal to 0.95. Now, let us ask ourselves the question, what is the rate of transmission of information? It is certainly less than 1000 bits/sec since on the average 50 out of every 1000 bits are incorrect. Our first impulse might be to say that the rate is 950 bits/sec by subtracting the number of errors from the data rate at the channel input. However, this is not satisfactory since the receiver does not know exactly which bits are in error, even though it knows that on the average 50 out of 1000 bits are incorrect. To further illustrate the difficulty in defining the amount of information transmitted as discussed above, let us consider the extreme case where the channel noise is so great that the probability of receiving a 1 or 0 is $\frac{1}{2}$ irrespective of what was transmitted. In such a case about $\frac{1}{2}$ of the received symbols are correct due to chance alone and we will give the system credit for transmitting 500 bits/sec, whereas no information is actually being transmitted. Indeed we can completely dispense with the channel and decide on the transmitted bit by flipping a coin at the receiving point, and correctly determine one half of the bits transmitted.

The inconsistency in defining information transmitted over a channel as the difference between input data rate and the error rate can be removed by making use of the information "lost" in the channel due to errors. In the preceding section we defined the conditional entropy of the input given the output $H(X|Y)$ as a measure of how uncertain we are of the input X given the output Y. For an ideal errorless channel we have no uncertainty about the input given the output and $H(X|Y)$ is equal to zero, that is, no information is lost. Knowing that $H(X|Y)=0$ for the ideal case wherein no information is lost, we may attempt to use $H(X|Y)$ to represent the amount of information lost in the channel. Accordingly we can define the amount of information transmitted over a channel by subtracting the information lost from the amount of information going into the channel. That is, we may define the *average rate of information transmission* D_t as

$$D_t \triangleq [H(X)-H(X|Y)]r_s \text{ bits/sec} \tag{4.32}$$

This definition takes care of the case when the channel is so noisy that the output is statistically independent of the input. When Y and X are independent, $H(X|Y)=H(X)$, and hence all the information going into the channel is lost and no information is transmitted over the channel. Let us illustrate these concepts by an example.

Example 4.7. A binary symmetric channel is shown in Figure 4.10. Find the rate of information transmission over this channel when $p=0.9$, 0.8, and 0.6; assume that the symbol (or bit) rate is 1000/sec.

Solution

$$H(X)=\tfrac{1}{2}\log_2 2+\tfrac{1}{2}\log_2 2 = 1 \text{ bit/symbol}$$
$$D_{in}=r_sH(X)=1000 \text{ bits/sec}$$

To find D_t, we need the conditional probabilities $P(X|Y)$, $X, Y=0, 1$. These conditional probabilities may be calculated as

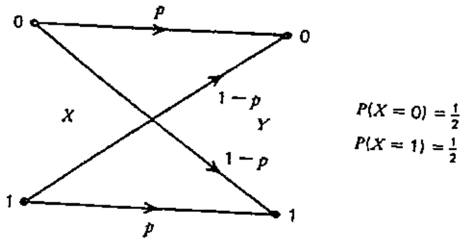


Figure 4.10 Binary symmetric channel.

$$P(X=0|Y=0)=\frac{P(Y=0|X=0)P(X=0)}{P(Y=0)}$$

and

$$P(Y=0)=P(Y=0|X=0)P(X=0)+P(Y=0|X=1)P(X=1)$$
$$=p(\tfrac{1}{2})+(1-p)\tfrac{1}{2}=\tfrac{1}{2}$$

Hence,

$$P(X=0|Y=0)=p$$

Similarly,

$$P(X=1|Y=0)=1-p$$
$$P(X=1|Y=1)=p$$
$$P(X=0|Y=1)=1-p$$

Hence,

$$\begin{aligned}H(X|Y)=&-P(X=0,\,Y=0)\log_2 P(X=0|Y=0)\\&-P(X=0,\,Y=1)\log_2 P(X=0|Y=1)\\&-P(X=1,\,Y=0)\log_2 P(X=1|Y=0)\\&-P(X=1,\,Y=1)\log_2 P(X=1|Y=1)\\=&-[\tfrac{1}{2}p\log_2 p+\tfrac{1}{2}(1-p)\log_2(1-p)\\&+\tfrac{1}{2}p\log_2 p+\tfrac{1}{2}(1-p)\log_2(1-p)]\\=&-[p\log_2 p+(1-p)\log_2(1-p)]\end{aligned}$$

and the rate of information transmission over the channel is given by

$$D_t=[H(X)-H(X|Y)]r_s \text{ bits/sec}$$

Values of D_t for $p=0.9$, 0.8, and 0.6 are given in Table 4.5.

The values shown in Table 4.5 clearly indicate that the rate of information transmission over the channel decreases very rapidly as the probability of error $1-p$ approaches $\frac{1}{2}$.

Table 4.5. Rate of information transmission versus values of p

p	0.9	0.8	0.6.
D_t	531 bits/sec	278 bits/sec	29 bits/sec

The reader should be aware of the fact that the *data rate* and *information rate* are two distinctly different quantities. With reference to Example 4.7, we often refer to the bit transition rate r_b at the channel input as the input data rate, or simply the bit rate. This is different from the information rate D_{in} at the channel input. D_{in} depends on r_b and the symbol (bit) probabilities. Furthermore, the rate of information transmission over the channel (D_t) depends not only on D_{in} but also on the channel symbol transition probabilities p_{ij}.

4.5.2 Capacity of a Discrete Memoryless Channel

The capacity of a noisy (discrete, memoryless) channel is defined as the maximum possible rate of information transmission over the channel. The maximum rate of transmission occurs when the source is "matched" to the channel. We define the *channel capacity* C as

$$\begin{aligned} C &\triangleq \max_{P(X)} \{D_t\} \\ &= \max_{P(X)} [H(X) - H(X|Y)]r_s \qquad (4.33) \end{aligned}$$

where the maximum is with respect to all possible information sources; that is, the maximum is taken with respect to all possible probability distributions for the discrete random variable X.

Example 4.8. Calculate the capacity of the discrete channel shown in Figure 4.11. Assume $r_s = 1$ symbol/sec.

Solution. Let $\alpha = -[p \log p + q \log q]$ and let $P(X=0) = P(X=3) = P$ and $P(X=1) = P(X=2) = Q$ (these probabilities being equal from consideration of symmetry). Then, from the definition of channel capacity,

$$C = \max_{P,Q} [H(X) - H(X|Y)]$$

subject to the constraint $2P + 2Q = 1$ (why?). From the definition of $H(X)$

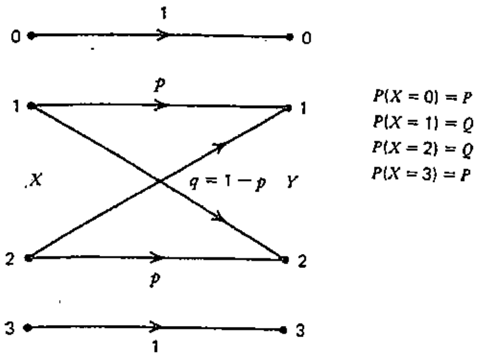


Figure 4.11 Channel model for Example 4.8.

and $H(X|Y)$, we obtain

$$\begin{aligned} H(X) &= -2P \log_2 P - 2Q \log_2 Q \\ H(X|Y) &= -2Q(p \log_2 p + q \log_2 q) = 2Q\alpha \end{aligned}$$

Hence,

$$D_t = -2P \log_2 P - 2Q \log_2 Q - 2Q\alpha$$

We want to maximize D_t with respect to P and Q, subject to $2P + 2Q = 1$ (or $Q = \frac{1}{2} - P$).

Substituting $Q = \frac{1}{2} - P$, we have

$$D_t = -2P \log_2 P - 2(\tfrac{1}{2} - P) \log_2(\tfrac{1}{2} - P) - 2(\tfrac{1}{2} - P)\alpha$$

To find the value of P that maximizes D_t, we set

$$\frac{dD_t}{dP} = 0$$

or

$$\begin{aligned} 0 &= -\log_2 e - \log_2 P + \log_2 e + \log_2(\tfrac{1}{2} - P) + \alpha \\ &= -\log_2 P + \log_2 Q + \alpha \end{aligned}$$

Solving for P, we get

$$\begin{aligned} P &= Q2^{\alpha} \\ &= Q\beta \end{aligned}$$

where

$$\beta = 2^{\alpha}$$

Substituting $P = Q\beta$ in $2P + 2Q = 1$, we can obtain the optimum values of

P and Q as

$$P = \frac{\beta}{2(1+\beta)}$$

$$Q = \frac{1}{2(1+\beta)}$$

The channel capacity is then,

$$\begin{aligned} C &= -2(P\log_2 P + Q\log_2 Q + Q\alpha) \\ &= -2\left[\frac{\beta}{2(1+\beta)}\log_2\left(\frac{\beta}{2(1+\beta)}\right) + \frac{1}{2(1+\beta)}\log_2\left(\frac{1}{2(1+\beta)}\right) + \frac{1}{2(1+\beta)}\log_2\beta\right] \\ &= \log_2\left(\frac{2(\beta+1)}{\beta}\right) \text{ bits/sec} \end{aligned}$$

A check with extreme values of $p = 1$ and $p = 0$ reveals the following: With $p = 1$, we have an errorless channel and the maximum rate of information transmission occurs when the input symbols occur with equal probability. The channel capacity for this ideal channel is 2 bits/symbol or 2 bits/sec with a symbol rate of 1 symbol/sec. For the noisy case with $p = \frac{1}{2}$, the capacity of the channel is $C = \log_2 3$. Here the first and fourth symbol are used more often than the other two because of their freedom from noise. Also the second and third symbols could not be distinguished at all and act together like one symbol. Hence, the capacity $\log_2 3$ seems to be a reasonable answer. For other values of p, the channel capacity will lie between $\log_2 3$ and $\log_2 4$ bits/sec.

The justification for defining a capacity for the noisy channel when we know that we can never send information without errors over such a channel is based on the fact that we can definitely reduce the probability of errors by repeating messages many times and studying the different received versions of the message. By increasing the redundancy of the encoding we can make the probability of error approach zero. This result is stated below as a theorem.

Theorem 4.2

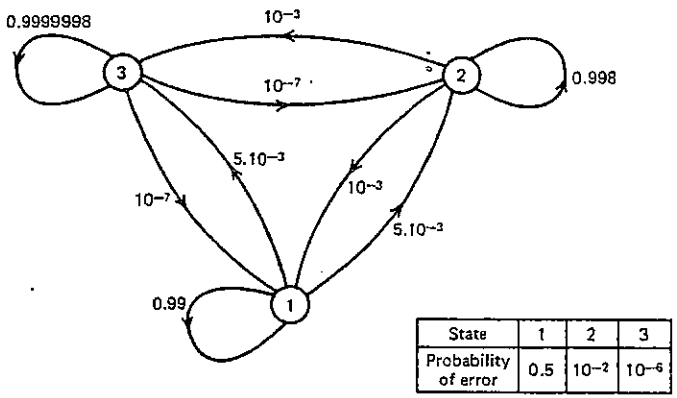
Let C be the capacity of a discrete memoryless channel, and let H be the entropy of a discrete information source emitting r_s symbols per second. If $r_s H \leq C$, then there exists a coding scheme such that the output of the source can be transmitted over the channel with an arbitrarily small probability of error. It is not possible to transmit information at a rate exceeding C without a positive frequency of errors.

While a proof of this theorem is mathematically formidable, we will look at encoding schemes that will accomplish the task mentioned in the theorem in a later chapter when we look at the design of channel encoders. For now, it suffices to say that if the information rate of the source is less than the capacity of the channel, then we can design a channel encoder/decoder that will allow us to transmit the output of the source over the channel with an arbitrarily small probability of error.

4.5.3 Discrete Channels with Memory

In the preceding sections we looked at channels that have no memory, that is, channels in which the occurrence of error during a particular symbol interval does not influence the occurrence of errors during succeeding symbol intervals. However, in many channels, errors do not occur as independent random events, but tend to occur in bursts. Such channels are said to have memory. Telephone channels that are affected by switching transients and dropouts, and microwave radio links that are subjected to fading are examples of channels with memory. In these channels, impulse noise occasionally dominates the Gaussian noise and errors occur in infrequent long bursts. Because of the complex physical phenomena involved, detailed characterization of channels with memory is very difficult.

A model that has been moderately successful in characterizing error bursts in channels is the Gilbert model. Here the channel is modeled as a discrete memoryless BSC, where the probability of error is a time varying parameter. The changes in probability of error is modeled by a Markoff process shown in Figure 4.12. The error generating mechanism in the channel occupies one of



State	1	2	3
Probability of error	0.5	10^{-2}	10^{-6}

Figure 4.12 A three-state Gilbert model for communication channels.

three states, and transition from one state to another is modeled by a discrete, stationary Markoff process. When the channel is in state 2 for example, bit error probability during a bit interval is 10^{-2} and the channel stays in this state during the succeeding bit interval with a probability 0.998. However, the channel may go to state 1 wherein the bit error probability is 0.5. Since the system stays in this state with probability 0.99, errors tend to occur in bursts (or groups). State 3 represents a low bit error rate, and errors in this state are produced by Gaussian noise. Errors very rarely occur in bursts while the channel is in this state. Other details of the model are shown in Figure 4.12. The maximum rate at which data can be transmitted over the channel can be computed for each state of the channel using the BSC model of the channel corresponding to each of the three states. Other characteristic parameters of the channel such as the mean time between error bursts, and mean duration of the error bursts can be calculated from the model.

4.6 CONTINUOUS CHANNELS

The communication channel between points d and f in Figure 4.7 is analog or continuous in nature. In this portion of the channel, the input signals are continuous functions of time, and the function of the channel is to produce at its output the electrical waveform presented at its input. A real channel accomplishes this only approximately. First, the channel modifies the waveform in a deterministic fashion, and this effect can be adequately modeled by treating the channel as a linear system. The channel also modifies the input waveform in a random fashion due to additive and multiplicative noise. Throughout this book we will deal with additive noise only since it occurs more often than multiplicative noise. Additive noise can be Gaussian or impulsive in nature. Gaussian noise includes thermal and shot noise from equipment and radiation picked up by the receiving antenna. According to the central limit theorem the noise that results from the summed effects of many sources tends to have a Gaussian distribution. Because of this omnipresence Gaussian noise is most often used to characterize the analog portion of communication channels. Modulation and demodulation techniques are designed with the primary objective of reducing the effects of Gaussian noise.

A second type of noise, impulse noise, is also encountered in the channel. Impulse noise is characterized by long quiet intervals followed by bursts of high amplitude noise pulses. This type of noise is due to switching transients, lightning discharges, and accidental hits during maintenance work, and so forth. The characterization of impulse noise is much more difficult than Gaussian noise. Also, analog modulation techniques are not as suitable as digital coding methods for dealing with impulse noise phenomena. For these reasons the effects of impulse noise are often included in the model of the discrete portion of the channel, and only Gaussian noise is included in the model of the analog portion of the channel.

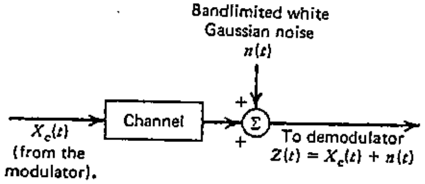

Bandlimited white Gaussian noise
$n(t)$
+
+
Channel
Σ
$X_c(t)$ (from the modulator).
To demodulator
$Z(t) = X_c(t) + n(t)$

Figure 4.13 Analog portion of the communication channel.

The analog portion of the communication channel can be modeled as shown in Figure 4.13 (see Section 3.7 and Figure 3.7). The input to the channel is a random process $X_c(t)$, which consists of the collection of all the waveforms generated by the modulator. The bandwidth of $X_c(t)$ and the channel is assumed to be B Hz (for convenience, let us assume $X_c(t)$ and the channel to be lowpass). The additive noise at the channel output is zero mean, bandlimited Gaussian white noise $n(t)$. The capacity of this portion of the channel is found by maximizing the rate of information transmission with respect to the distribution of $X_c(t)$. While the formulation of this problem is similar to the one we used for discrete channels in terms of $H(X_c)$ and $(H(X_c/Z)$, the optimization is very involved. However, the result has a very simple form; we state the result as a theorem (for a direct proof see Shannon's book*) and discuss how the result can be used in the design of communication systems.

4.6.1 Shannon–Hartley Theorem and Its Implications

Theorem 4.3

The capacity of a channel with bandwidth B and additive Gaussian bandlimited white noise is

$$C = B \log_2(1 + S/N) \text{ bits/sec} \tag{4.34}$$

where S and N are the average signal power and noise power, respectively, at the output of the channel. ($N = \eta B$ if the two-sided power spectral density of the noise is $\eta/2$ watts/Hz.)

*We give an indirect proof of Shannon's theorem in Section 8.7.3. Also, see Problem 4.27 in which the reader is asked to derive a relationship similar to the one given in Equation (4.34).

This theorem, referred to as the Shannon–Hartley theorem, is of fundamental importance and has two important implications for communication systems engineers. First, it gives us the upper limit that can be reached in the way of reliable data transmission rate over Gaussian channels. Thus a system designer always tries to optimize his system to have a data rate as close to C given in Equation (4.34) as possible with an acceptable error rate.

The second implication of the Shannon–Hartley theorem has to do with the exchange of signal-to-noise ratio for bandwidth. To illustrate this aspect of the theorem, suppose that we want to transmit data at a rate of 10,000 bits/sec over a channel having a bandwidth $B = 3000$ Hz. To transmit data at a rate of 10,000 bits/sec, we need a channel with a capacity of at least 10,000 bits/sec. If the channel capacity is less than the data rate, then errorless transmission is not possible. So, with $C = 10{,}000$ bits/sec we can obtain the (S/N) requirement of the channel as

$$\begin{aligned}(S/N) &= 2^{(C/B)} - 1\\ &= 2^{3.333} - 1 \approx 9\end{aligned}$$

For the same problem if we have a channel with a bandwidth of 10,000 Hz, then we need a S/N ratio of 1. Thus a bandwidth reduction from 10,000 Hz to 3000 Hz results in an increase in signal power from 1 to 9.

Another interesting aspect of the Shannon–Hartley theorem has to do with *bandwidth compression*. To illustrate this aspect let us ask ourselves the question, is it possible to quantize and transmit a signal whose spectral range extends up to a frequency f_m over a channel having a bandwidth less than f_m? The answer is yes and we can justify the answer as follows. Suppose we sample the analog signal at a rate of $3f_m$ samples/sec (i.e., at 1.5 times the Nyquist rate for example) and quantize the signal value into one of M possible levels. Then the data rate of the quantized signal is $3f_m \log_2 M$ bits/sec. If the bandwidth of the channel is B, then by an appropriate choice of signal power we can achieve a capacity C greater than $3f_m \log_2 M$. For example, with $M = 64$ and a channel bandwidth equal to half of the signal bandwidth, we would need a S/N ratio of about 109 dB to be able to transmit the quantized signal with a small probability of error. Thus a bandwidth compression by a factor of 2 is possible if we can maintain a S/N ratio of 109 dB (an impractical value) at the output of the channel. We are assuming that signal distortion due to sampling and quantizing is negligible.

The Shannon–Hartley theorem indicates that a *noiseless channel has an infinite capacity*. However, when noise is present the channel capacity does not approach infinity as the bandwidth is increased because the noise power increases as the bandwidth increases. The channel capacity reaches a *finite* upper limit with *increasing bandwidth* if the signal power is fixed. We can calculate this limit as follows. With $N = \eta B$, where $\eta/2$ is the noise power spectral density, we have

$$\begin{aligned}C &= B\log_2\left(1+\frac{S}{\eta B}\right)\\ &= \left(\frac{S}{\eta}\right)\left(\frac{\eta B}{S}\right)\log_2\left(1+\frac{S}{\eta B}\right)\\ &= \left(\frac{S}{\eta}\right)\log_2\left(1+\frac{S}{\eta B}\right)^{\eta B/S} \qquad (4.35)\end{aligned}$$

Recalling that $\lim_{x\to 0}(1+x)^{1/x} = e$ and letting $x = S/\eta B$ in (4.35), we have

$$\begin{aligned}\lim_{B\to\infty} C &= \frac{S}{\eta}\log_2 e\\ &= 1.44\left(\frac{S}{\eta}\right) \qquad (4.36)\end{aligned}$$

A communication system capable of transmitting information at a rate of $B\log_2(1+S/N)$ is called an ideal system. Shannon proposed the following idea for such a system. Let us assume that the source puts out M equiprobable messages of length T seconds. The ideal communication system observes the source output for T seconds and the message is represented (encoded) by a channel signal chosen from a collection of M sample functions of white noise of duration T. At the output of the channel, the received signal plus noise is compared with stored versions of the channel signals. The channel signal that "best matches" the signal plus noise is presumed to have been transmitted and the corresponding message is decoded. The total amount of time delay involved in observing the message signal, transmitting, and decoding at the receiver is at best T seconds.

The ideal signalling scheme using noiselike signals can convey information at a rate approaching the channel capacity only when $T\to\infty$. Only in the limiting case do we have all the conditions satisfied. Under this limiting condition, the ideal system has the following characteristics:

1. The information rate → $B\log_2(1+S/N)$.
2. The error rate → 0.
3. The transmitted and received signals have the characteristics of bandlimited Gaussian white noise.
4. As $T\to\infty$, the number of signals $M\to\infty$ and the coding delay also → ∞.

It must be obvious from the preceding discussion that an ideal system cannot be realized in practice. Rather than trying to design a system using a large number of analog signals, we use a small number of analog signals in practical systems. This leads to a nonzero probability of error P_e. The data rate and the error probability define a discrete channel whose capacity C' will be less than $B\log_2(1+S/N)$. Through this digital channel we try to achieve a data rate approaching C' with a probability of error approaching zero using digital error control encoding. Thus in practical systems we seldom try to achieve the maximum theoretical rate of information transmission over the analog portion of the channel. We keep this portion of the system reasonably simple. In the digital portion of the system, we try to achieve a rate approaching the capacity of the discrete portion of the channel since digital encoding is easier to implement.

In the following chapters we will discuss signaling schemes for transmitting digital information through an analog communication channel. For each type of signaling scheme, we will derive expressions for the error probability in terms of the bandwidth required, output S/N, and the data rate. These relationships define the parameters of the discrete portion of the channel. In Chapter 9 we will look at methods of error control coding that will enable us to transmit information over the discrete channel at a rate approaching its capacity with a small probability of error.

Before we conclude our discussion of the Shannon–Hartley theorem, it must be pointed out that the result given in Equation (4.34) is for the Gaussian channel. This limitation does not in any way diminish the importance and usefulness of the Shannon–Hartley law for the following reasons: First, most physical channels are generally at least approximately Gaussian. Second, it has been shown that the result obtained for the Gaussian channel provides a *lower bound* on the performance of a system operating over a non-Gaussian channel. That is, if a particular encoder/decoder yields an error probability P_e over the Gaussian channel, another encoder/decoder can be designed for a non-Gaussian channel to yield a smaller probability of error. Detailed expressions for the channel capacity have been derived for several non-Gaussian channels.

We now present several examples that illustrate the use of channel capacity in the design of communication systems.

Example 4.9. Calculate the capacity of a lowpass channel with a usable bandwidth of 3000 Hz and $S/N = 10^3$ at the channel output. Assume the channel noise to be Gaussian and white.

Solution. The capacity C is given by Equation (4.34) as

$$C = B\log_2\left(1+\frac{S}{N}\right)$$
$$= (3000)\log_2(1+1000)$$
$$\approx 30{,}000 \text{ bits/sec}$$

The parameter values used in this example are typical of standard voice grade telephone lines. The maximum data rate achievable now on these channels is 9600 bits/sec. Rates higher than this require very complex modulation and demodulation schemes.

Example 4.10. An ideal lowpass channel of bandwidth B Hz with additive Gaussian white noise is used for transmitting digital information. (a) Plot C/B versus (S/N) in dB for an ideal system using this channel. (b) A practical signaling scheme on this channel uses one of two waveforms of duration T_b seconds to transmit binary information. The signaling scheme transmits data at a rate of $2B$ bits/sec, and the probability of error is given by*

$$P(\text{error}|1\text{ sent}) = P(\text{error}|0\text{ sent}) = P_e$$
$$= Q(\sqrt{S/N})$$

where

$$Q(z) = \int_z^\infty \frac{1}{\sqrt{2\pi}}\exp(-x^2/2)\,dx$$

Using the tabulated values of $Q(z)$ given in Appendix D, plot the rate of information transmission versus (S/N) in dB for this scheme.

Solution

(a) For the ideal scheme, we have

$$\frac{C}{B} = \log_2\left(1+\frac{S}{N}\right)$$

When $S/N \gg 1$, $C/B \approx \log_2(S/N)$.

(b) The binary signaling scheme corresponds to a discrete binary symmetric channel with $P_e = Q(\sqrt{S/N})$. The information rate over this channel is giver by (see Example 4.7)

$$D_t = 2B[1 - P_e\log_2(1/P_e) - (1-P_e)\log_2(1/1-P_e)]$$
$$= 2B[1 + P_e\log_2 P_e + (1-P_e)\log_2(1-P_e)]$$

For large S/N ratios, $P_e \approx 0$ and $D_t = 2B$. It also represents the capacity or the maximum rate at which we can transmit information using the binary signaling scheme. Values of C/B and D_t/B for various values of S/N are shown in Table 4.6 and shown plotted in Figure 4.14. These results show that for high S/N ratios, the binary signaling scheme is very inefficient. With high S/N ratios we can transmit and decode correctly a large

*Expressions for probabilities of error for various signaling schemes will be derived in Chapters 5 and 8.

Table 4.6. C/B and D_t/B for various values of S/N

$(S/N)_{\text{dB}}$	0	6	10	20
$(C/B)_{\text{ideal}}$	1	2.32	3.46	6.65
P_e	0.159	0.028	0.0008	≈ 0
$(D_t/B)_{\text{binary}}$	0.7236	1.6314	1.98	2.0

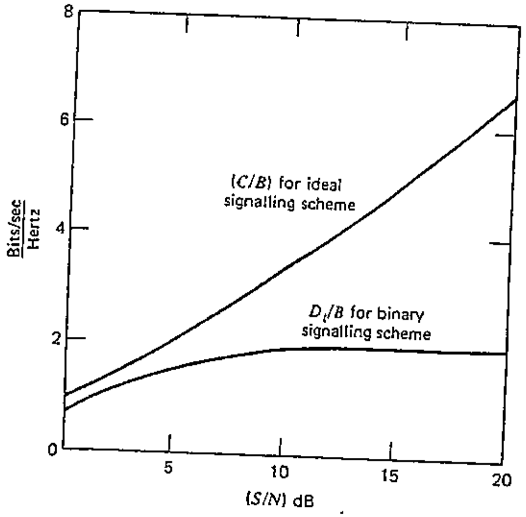


Figure 4.14 Plots of C/B and D_t/B.

number of waveforms and hence an M-ary signaling scheme, $M > 2$ should be used (we will discuss M-ary signaling schemes in Chapters 5 and 8).

The following example illustrates how we can use the concept of information rate and channel capacity in the design of communication systems.

Example 4.11. A CRT terminal is used to enter alphanumeric data into a computer. The CRT is connected to the computer through a voice grade telephone line having a usable bandwidth of 3000 Hz and an output S/N of 10 dB. Assume that the terminal has 128 characters and that the data sent from the terminal consist of independent sequences of equiprobable characters.

(a) Find the capacity of the channel.

(b) Find the maximum (theoretical) rate at which data can be transmitted from the terminal to the computer without errors.

Solution

(a) The capacity is given by:

$$C = B \log_2(1 + S/N)$$
$$= (3000) \log_2(11) = 10{,}378 \text{ bits/sec}$$

(b) Average information content/character:

$$H = \log_2(128) = 7 \text{ bits/character}$$

and the average information rate of the source $R = r_s H$. For errorless transmission, we need $R = r_s H < C$ or

$$7r_s < 10{,}378$$
$$r_s < 1482$$

Hence the maximum rate at which data can be transmitted without errors is 1482 characters/sec.

4.7 SUMMARY

A probabilistic model for discrete information sources was developed and the entropy of the source and the average information rate of the source were defined. The source entropy has the units of bits per symbol and it represents the average number of bits per symbol needed to encode long sequences of symbols emitted by the source. The average information rate represents the average number of bits per second needed to encode the source output. The functional block that maps the symbol sequence emitted by the source into a binary data stream is the source encoder. A procedure for designing an encoder using the algorithm given by Shannon was presented. The effect of design parameters such as block length and code word lengths on the complexity of the encoder, time delay in decoding and the efficiency of the encoder were discussed.

Mathematical models for discrete and continuous channels were discussed. The capacity of a channel represents the maximum rate at which data can be transmitted over the channel with an arbitrarily small probability of error. It was pointed out that the maximum rate of data transmission over a channel can be accomplished only by using signals of large dimensionality.

Several examples were presented to illustrate the concepts involved in modeling and analyzing discrete information sources and communication channels.

REFERENCES

Most of the material presented in this chapter is based on Shannon's work, which was first published in 1948. This classical work is very readable and quite interesting. Treatment of concepts in information theory and coding theory at an advanced level may be found in the books of Abramson (1963), and Wozencraft and Jacobs (1965) and Gallager (1968). Undergraduate texts such as the ones by Carlson (1975), and Taub and Schilling (1971) provide a brief treatment of information theory.

In order to understand and appreciate the concepts of source and channel models, the reader should be familiar with the theory of random variables and random processes. Books by Breipohl (1970—written for undergraduates) and by Papoulis (1965—beginning graduate level) provide thorough and easily readable treatment of the theory of random variables and random processes. Topics in measurement and analysis of random data are treated well in the book by Bendat and Piersol (1971).

1. C. E. Shannon. *Mathematical Theory of Communication.* Univ. of Illinois Press (1963) (original work was published in the Bell System Technical Journal, vol. 27, 379–423, 623–656, 1948).
2. N. Abramson. *Information Theory and Coding.* McGraw-Hill, New York (1963).
3. J. M. Wozencraft and I. M. Jacobs. *Principles of Communication Engineering.* Wiley, New York (1965).
4. R. G. Gallager. *Information Theory and Reliable Communication.* Wiley, New York (1968).
5. H. Taub and D. L. Schilling. *Principles of Communication Systems.* McGraw-Hill, New York (1971).
6. A. B. Carlson. *Communication Systems.* McGraw-Hill, New York (1975).
7. A. M. Breipohl. *Probabilistic Systems Analysis.* Wiley, New York (1970).
8. A. Papoulis. *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill, New York (1965).
9. J. S. Bendat and A. G. Piersol. *Measurement and Analysis of Random Data.* Wiley, New York (1971).

PROBLEMS

Section 4.2

✓ 4.1. A source emits one of four possible messages m_1, m_2, m_3, and m_4 with probabilities $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, and $\frac{1}{8}$, respectively. Calculate the information content of each message and the average information content per message.

4.2. A card is drawn from a deck of playing cards. (a) You the card you draw is a spade. How much information c bits)? (b) How much information do you receive if you card that you drew is an ace? (c) How much informatio if you are told that the card you drew is an ace o information content of the message "ace of spades" information contents of the messages "spade" and "ac

✓ 4.3. A source emits an independent sequence of symbols f consisting of five symbols A, B, C, D, and E with symbo $\frac{1}{8}$, $\frac{1}{8}$, $\frac{3}{16}$, $\frac{5}{16}$, respectively. Find the entropy of the source.

✓ 4.4. A binary source is emitting an independent sequence of probabilities p and $1-p$, respectively. Plot the entropy o versus p $(0 < p < 1)$.

✓ 4.5. For a source emitting symbols in independent sequences, show th source entropy is maximum when the symbols occur with equal probabilities.

4.6. The international Morse code uses a sequence of dots and dashes to transmit letters of the English alphabet. The dash is represented by a current pulse that has a duration of 3 units and the dot has a duration of 1 unit. The probability of occurrence of a dash is $\frac{1}{3}$ of the probability of occurrence of a dot.

(a) Calculate the information content of a dot and a dash.
(b) Calculate the average information in the dot–dash code.
(c) Assume that the dot lasts 1 msec, which is the same time interval as the pause between symbols. Find the average rate of information transmission.

✓ 4.7. The probability of occurrence of the various letters of the English alphabet are given below:

A	0.081	J	0.001	S	0.066
B	0.016	K	0.005	T	0.096
C	0.032	L	0.040	U	0.031
D	0.037	M	0.022	V	0.009
E	0.124	N	0.072	W	0.020
F	0.023	O	0.079	X	0.002
G	0.016	P	0.023	Y	0.019
H	0.051	Q	0.002	Z	0.001
I	0.072	R	0.060		

(a) What letter conveys the maximum amount of information?

(b) What letter conveys the minimum amount of information?
(c) What is the entropy of English text if you can assume that letters are chosen independently to form words and sentences (not a realistic assumption!).
(d) If I am thinking of a word and tell you the first letter of the word, which will be a more helpful clue, T or X? Why?

✓ 4.8. A black and white TV picture consists of 525 lines of picture information. Assume that each line consists of 525 picture elements and that each element can have 256 brightness levels. Pictures are repeated at the rate of 30/sec. Calculate the average rate of information conveyed by a TV set to a viewer.

✓ 4.9. The output of an information source consists of 128 symbols, 16 of which occur with a probability of 1/32 and the remaining 112 occur with a probability of 1/224. The source emits 1000 symbols/sec. Assuming that the symbols are chosen independently, find the average information rate of this source.

✓ 4.10. The state diagram of a stationary Markoff source is shown in Figure 4.15.
(a) Find the entropy of each state H_i ($i = 1, 2, 3$).
(b) Find the entropy of the source H.
(c) Find G_1, G_2, and G_3 and verify that $G_1 \geq G_2 \geq G_3 \geq H$.

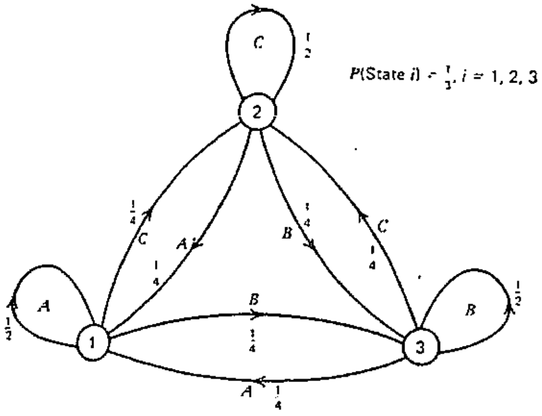


Figure 4.15 Source diagram for Problem 4.10.

4.11. Re-work the previous problem for the source shown in Figure 4.3.

Section 4.3

4.12. For the source described in Example 4.5:
(a) Design a source encoding scheme using a block size of two symbols and variable length code words. Calculate the actual number of bits per symbol $\hat{H}_2$ used by the encoder and verify that

$$\hat{H}_2 \leq G_2 + \tfrac{1}{2}$$

(b) Design a source encoding scheme using fixed length code words and a block size of four symbols. Compute the actual number of bits per symbol used.
(c) If the source is emitting symbols at a rate of 1000 symbols/sec, compute the output bit rate of the encoders (a) and (b).

4.13. Another technique used in constructing a source encoder consists of arranging the messages in decreasing order of probability and dividing the message into two almost equally probable groups. The messages in the first group are given the bit 0 and the messages in the second group are given the bit 1. The procedure is now applied again for each group separately, and continued until no further division is possible. Using this algorithm, find the code words for six messages occurring with probabilities $\frac{1}{3}, \frac{1}{3}, \frac{1}{6}, \frac{1}{12}, \frac{1}{24}, \frac{1}{24}$.

4.14. Another way of generating binary code words for messages consists of arranging the messages in decreasing order of probability and dividing code words as follows: The code word for the first message is "0". The code word for the ith message consists of $(i-1)$ bits of "1's" followed by a "0." The code word for the last message consists of all "1's"; the number of bits in the code word for the last message is equal to the total number of messages that are to be encoded.
(a) Find the code words and the average number of bits per message used if the source emits one of five messages with probabilities $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}$, and $\frac{1}{16}$.
(b) Is this code uniquely decipherable? That is, for every possible sequence of bits, is there only one way of interpreting the messages?

4.15. A source emits *independent* sequences of symbols from a source alphabet containing five symbols with probabilities 0.4, 0.2, 0.2, 0.1 and 0.1.
(a) Compute the entropy of the source.
(b) Design a source encoder with a block size $n = 2$.

Section 4.5

✓ 4.16. A nonsymmetric binary channel is shown in Figure 4.16.

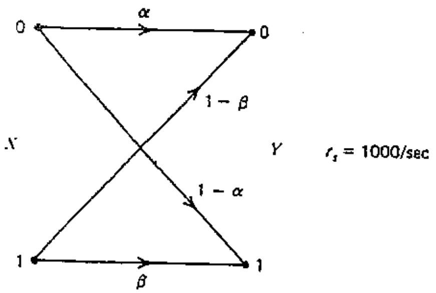


Figure 4.16 Binary channel model for Problem 4.17.

(a) Find $H(X)$, $H(Y)$, $H(X|Y)$, and $H(Y|X)$ when $P(X=0)=\frac{1}{4}$, $P(X=1)=\frac{3}{4}$, $\alpha=0.75$, and $\beta=0.9$.

(b) Find the capacity of the channel for $\alpha=0.75$ and $\beta=0.9$.

(c) Find the capacity of the binary symmetric channel ($\alpha=\beta$).

4.17. Show that $H(X, Y)=H(X)+H(Y|X)=H(Y)+H(X|Y)$.

4.18. Find the capacity of the discrete channel shown in Figure 4.17.

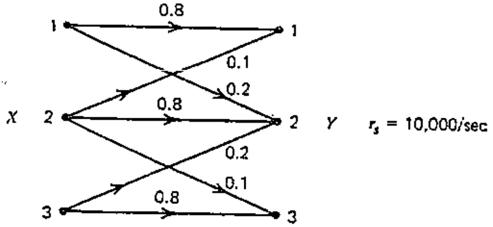


Figure 4.17 Channel model for Problem 4.18.

✓4.19. Show that (a) $H(X|Y)=H(X)$ when X and Y are statistically independent, and (b) $H(X|Y)=0$ when $X=Y$.

✓4.20. A discrete channel accepts as its input a binary sequence with a bit rate of r_b bits/sec. The channel signals are selected from a set of eight possible waveforms, each having a duration $3/r_b$ seconds. Thus, each waveform may convey up to three bits of information. The channel noise is such that when the received waveform is decoded, each block of three input bits is received with no errors, or with exactly one error in the first, second, or third bit position. Assuming that these four outcomes are equally likely to occur:

(a) Find the capacity of the discrete channel.

(b) Suppose that you want to transmit the output of an information source having a rate $R=r_b/3$ over this channel. How would you encode the data so that errorless transmission is possible?

4.21. The state model of a discrete channel with memory is shown in Figure 4.18. In state 1, the channel corresponds to a BSC with an error-probability of 0.001. At state 2, the channel is again a BSC with an error probability of 0.5. The state and transitional probabilities are shown in the diagram. Assume that the bit rate at the input to the channel is 1000 bits/sec and the transition rate of the state of the channel is also 1000/sec.

(a) Find the capacity of the channel for state 1 and state 2.

(b) Find the average capacity of the channel.

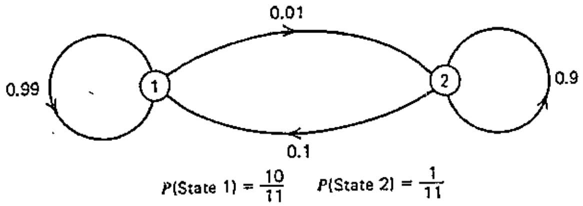

0.01
0.99
1
2
0.9
0.1
$P(\text{State 1}) = \frac{10}{11}$
$P(\text{State 2}) = \frac{1}{11}$

Figure 4.18 Channel state model for Problem 4.12.

Section 4.6

✓4.22. Calculate the capacity of a Gaussian channel with a bandwidth of 1 MHz and S/N ratio of 30 dB.

4.23. How long will it take to transmit one million ASCII characters over the channel in Problem 4.22? (In ASCII code, each character is coded as an 8-bit binary word; ignore start and stop bits.)

✓4.24. A Gaussian channel has a bandwidth of 4 kHz and a two-sided noise power spectral density $\eta/2$ of 10^{-14} watt/Hz. The signal power at the receiver has to be maintained at a level less than or equal to 1/10 of a milliwatt. Calculate the capacity of this channel.

✓4.25. An analog signal has a 4 kHz bandwidth. The signal is sampled at 2.5 times the Nyquist rate and each sample is quantized into one of 256 equally likely levels. Assume that the successive samples are statistically independent.

(a) What is the information rate of this source?

(b) Can the output of this source be transmitted without errors over a Gaussian channel with a bandwidth of 50 kHz and S/N ratio of 23 dB?

(c) What will be the bandwidth requirements of an analog channel for transmitting the output of the source without errors if the S/N ratio is 10 dB?

4.26. A friend of yours says that he can design a system for transmitting the output of a minicomputer to a line printer operating at a speed of 30 lines/minute over a voice grade telephone line with a bandwidth of 3.5 kHz, and $S/N = 30$ dB. Assume that the line printer needs eight bits of data per character and prints out 80 characters per line. Would you believe him?

✓ 4.27. The waveform shown in Figure 4.19 is used for transmitting digital information over a channel having a bandwidth $B \approx 1/2T$. Assume that the M levels are equally likely to occur and that they occur as an independent sequence.

(a) Find $E\{X^2(t)\}$.

(b) Find S/N (note: $N = \sigma^2$, and $S = E\{X^2(t)\}$).

(c) Assume λ is large enough so that errors occur with a probability $P_e \to 0$. Find the rate of information conveyed by the signal. Compare your result with Equation (4.34).

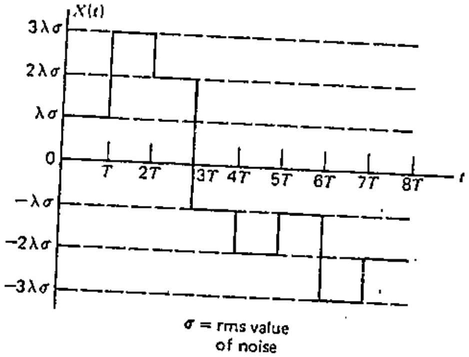


Figure 4.19 Signal waveform; sequences of levels convey messages.

5

BASEBAND DATA TRANSMISSION

5.1 INTRODUCTION

In the previous chapter, we discussed the theoretical limitations on the rate of information transmission over noisy channels. We pointed out that the maximum errorless rate of data transmission over a noisy channel could be achieved only by using signal sets of large dimensionality whose statistical characteristics match the noise characteristics. In a practical system, the large dimensionality of signals is realized by the digital encoding of a small set of basic waveforms generated by the modulator. The number of analog waveforms generated by commercial digital modulators range from 2 (binary) to a maximum of 8 or 16. There are many types of modulators corresponding to the many possible selections of modulator waveforms. In this chapter we take a detailed look at the analysis and design of discrete pulse modulation techniques that can be used for transmitting the output of a discrete source over a *baseband* channel. In a later chapter we will discuss discrete carrier modulation schemes that are used for transmitting digital information over *bandpass* channels.

In discrete pulse modulation, the amplitude, duration or position of the transmitted pulses is varied according to the digital information to be transmitted. These pulse modulation schemes are referred to as pulse amplitude (PAM), pulse duration (PDM), and pulse position modulation (PPM) schemes.

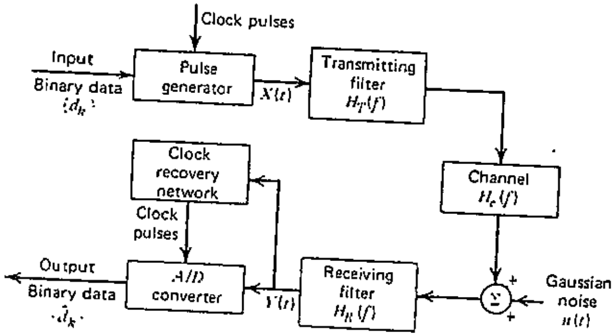


Figure 5.1 Baseband binary data transmission system.

Of these three methods, PAM systems are most efficient in terms of power and bandwidth utilization. This chapter is devoted to the study of PAM systems.

The elements of a baseband binary PAM system are shown in Figure 5.1. The input to the system is a binary data sequence with a bit rate of r_b and bit duration of T_b. The pulse generator output is a pulse waveform

$$X(t) = \sum_{k=-\infty}^{\infty} a_k p_g(t - kT_b) \tag{5.1a}$$

where $p_g(t)$ is the basic pulse whose amplitude a_k depends on the kth input bit. For convenience we will assume that $p_g(t)$ is normalized such that

$$p_g(0) = 1 \tag{5.1b}$$

and

$$a_k = \begin{cases} a & \text{if } k\text{th input bit is 1} \\ -a & \text{if } k\text{th input bit is 0} \end{cases} \tag{5.1c}$$

The PAM signal $X(t)$ passes through a transmitting filter $H_T(f)$, and then through the channel that adds random noise in addition to modifying the signal in a deterministic fashion. The noisy signal then goes through the receiving filter $H_R(f)$, and the output $Y(t)$ of the receiving filter is sampled by the analog-to-digital (A/D) converter. The transmitted bit stream is regenerated by the A/D converter based on the sampled values of $Y(t)$. The sampling instant is determined by the clock or timing signal that is usually generated from $Y(t)$ itself. A set of typical waveforms that occur at various points in the system is shown in Figure 5.2.

The A/D converter input $Y(t)$ can be written as

$$Y(t) = \sum_k A_k p_r(t - t_d - kT_b) + n_0(t) \tag{5.1d}$$

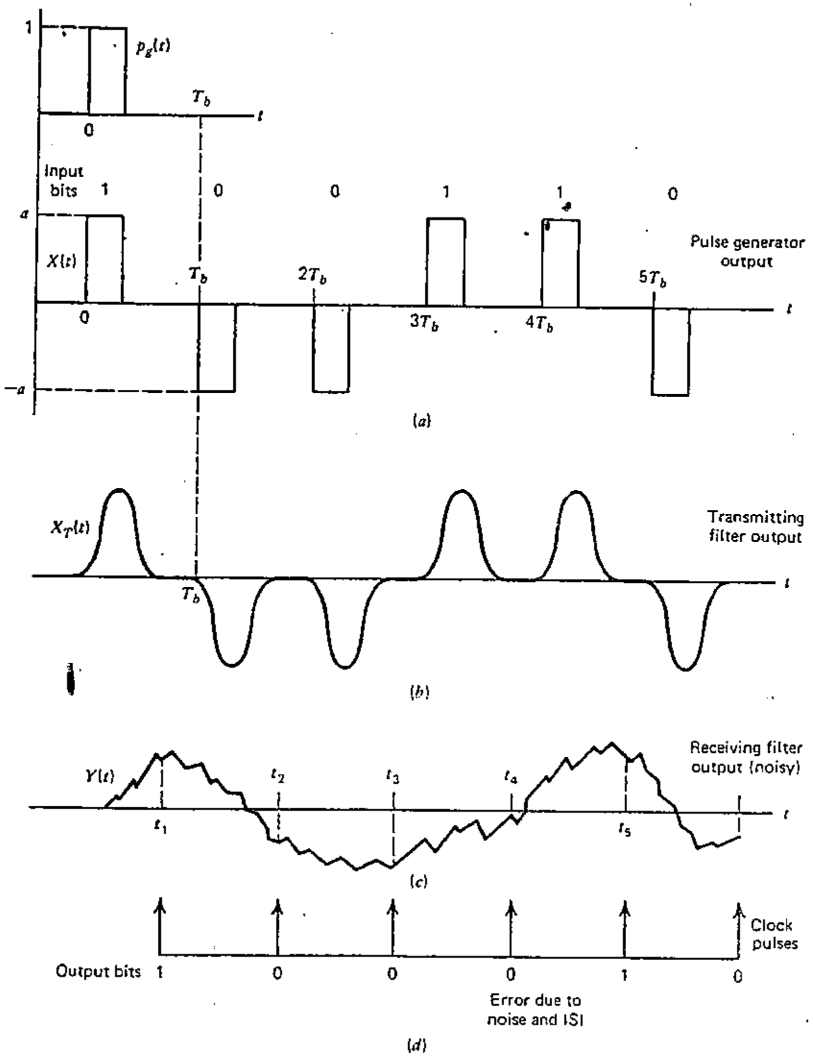


Figure 5.2 Example of typical waveforms in a binary PAM system. (*a*) Pulse generator output. (*b*) Transmitting filter output. (*c*) Receiving filter output (noisy). (*d*) Clock pulses.

where $A_k = K_c a_k$ and $K_c p_r(t - t_d)$ is the response of the system when the input is $p_g(t)$. In Equation (5.1d), t_d is an arbitrary time delay and $n_0(t)$ is the noise at the receiver output. K_c is a normalizing constant that yields $p_r(0) = 1$.

The A/D converter samples $Y(t)$ at $t_m = mT_b + t_d$ and the mth output bit is generated by comparing $Y(t_m)$ with a threshold (which is 0 for a symmetric binary PAM data transmission system). The input to the A/D converter at the sampling instant $t_m = mT_b + t_d$ is

$$Y(t_m) = A_m + \sum_{k \neq m} A_k p_r[(m - k)T_b] + n_0(t_m) \tag{5.2}$$

In Equation (5.2), the first term A_m represents the mth transmitted bit while the second term represents the باقیمانده residual effect of all other transmitted bits on the mth bit being decoded. This residual effect is called the *intersymbol interference* (ISI). The last term in (5.2) represents the noise.

In the absence of noise, and in the absence of ISI, the mth transmitted bit can be decoded correctly based on $Y(t_m)$ since $Y(t_m) = K_c a_m$ and a_m is uniquely related to the mth input bit. Noise and ISI introduce errors in the output. The major objectives of baseband PAM system design are to choose the transmitting and receiving filters to minimize the effects of noise and eliminate or minimize ISI. In addition, for a given transmitted power it may be مطلوب desirable to maximize the signaling rate r_b for a given bandwidth B or minimize the bandwidth required for a given signaling rate.

We deal with the design and analysis of optimum baseband PAM system in this chapter. The design of the system starts with the assumption that the physical characteristics of the channel and the statistical characteristics of the noise and the input bit stream are known. The pulse shapes $p_g(t)$ and $p_r(t)$ and the transfer functions of the filters $H_T(f)$ and $H_R(f)$ are to be chosen to optimize the performance of the system, keeping the bit error probability below a specified value. The bit error probability, which is commonly used as a measure of performance for binary PAM systems, is defined as $P_e = P[\hat{d}_k \neq d_k]$ (see Figure 5.1).

The design and analysis of a binary baseband PAM system is treated first and the M-ary PAM schemes are treated next. We will see that the M-ary schemes require a smaller bandwidth but more power than a binary scheme for a given data rate and bit error probability. Special signaling schemes, such as the duobinary scheme, are discussed and the effects of precoding the input bit stream on the spectral characteristics of the transmitted signal are illustrated.

The design معیارها criteria are aimed at an overall pulse shaping that would yield zero ISI. However, in practical systems some amount of residual ISI will [illegible] inevitably occur due to imperfect filter realization, incomplete knowledge of channel characteristics, and changes in channel characteristics. Hence, an equalizing filter is often inserted between the receiving filter and the A/D converter to compensate for changes in the parameters of the channel. We will look at procedures for designing zero forcing equalizers that will reduce the ISI.

In the PAM method of data transmission a clock signal must be recovered at the receiving end to set the sampling rate and sampling times. The clock information must somehow be carried in the transmitted signal. Methods of carrying the clock information and recovering it vary with the pulse shapes and coding methods used. We will discuss several methods of clock recovery at the receiver.

It must be pointed out that the important parameters involved in the design of a PAM system are data rate, error rate, transmitted power, noise power spectral density, and system complexity. These parameters are interrelated and the design procedure will involve trade-offs between the parameters to arrive at a system that meets the specified performance requirements and the given constraints.

5.2 BASEBAND BINARY PAM SYSTEMS

In this section we deal with the design of optimum baseband binary data transmission systems. Data rates in binary systems may range from a low rate of 100 bits/sec (BPS) in applications involving electromechanical devices such as a teletype to a high rate of up to tens of megabits per second in applications involving data transfer between computers. The rate typically is from 300 to 4800 bits/sec over voice grade telephone links to several hundred megabits per second over wideband microwave radio links. The acceptable bit error rate varies over a wide range depending on the application. Error probabilities in the range of 10^{-4} to 10^{-6} are representative and suitable to many applications. For design purposes we will assume that the input data rate and overall bit error probability are specified. Furthermore, we will assume that the characteristics of the channel are given, and that the channel noise can be represented by a zero mean Gaussian random process with a known power spectral density $G_n(f)$. The source that generates the input bit stream will be assumed to be ergodic and the source output will be assumed to be independent sequences of equiprobable bits.

The design of a baseband binary PAM system consists of specifying the pulse shapes $p_g(t)$ and $p_r(t)$ and the filters $H_R(f)$ and $H_T(f)$ to minimize the combined effects of intersymbol interference and noise in order to achieve a minimum probability of error for given data rate and power levels in the system.

5.2.1 Baseband Pulse Shaping

The intersymbol interference given by the second term in Equation (5.2) can be eliminated by proper choice of the received pulse shape $p_r(t)$. An inspection of Equation (5.2) reveals that for zero ISI, $p_r(t)$ should satisfy

$$p_r(nT_b) = \begin{cases} 1 & \text{for } n = 0 \\ 0 & \text{for } n \neq 0 \end{cases} \tag{5.3}$$

The constraint stated in (5.3) does not uniquely specify $p_r(t)$ for all values of t. To meet the constraint given in Equation (5.3), the Fourier transform $P_r(f)$ of $p_r(t)$ needs to satisfy a simple condition stated below.

Theorem 5.1

If $P_r(f)$ satisfies

$$\sum_{k=-\infty}^{\infty} P_r\left(f + \frac{k}{T_b}\right) = T_b \quad \text{for } |f| < 1/2T_b \tag{5.4}$$

then

$$p_r(nT_b) = \begin{cases} 1 & \text{for } n = 0 \\ 0 & \text{for } n \neq 0 \end{cases} \tag{5.5}$$

Proof

$p_r(t)$ is related to $P_r(f)$ by

$$p_r(t) = \int_{-\infty}^{\infty} P_r(f) \exp(j2\pi ft)\, df$$

The range of integration in the preceding equation can be divided into segments of length $1/T_b$ as

$$p_r(t) = \sum_{k=-\infty}^{\infty} \int_{(2k-1)/2T_b}^{(2k+1)/2T_b} P_r(f) \exp(j2\pi ft)\, df$$

and we can write $p_r(nT_b)$ as

$$p_r(nT_b) = \sum_k \int_{(2k-1)/2T_b}^{(2k+1)/2T_b} P_r(f) \exp(j2\pi fnT_b)\, df$$

Making a change of variable, $f' = f - k/T_b$, we can write the above equation as

$$p_r(nT_b) = \sum_k \int_{-1/2T_b}^{1/2T_b} P_r\left(f' + \frac{k}{T_b}\right) \exp(j2\pi f'nT_b)\, df'$$

Further, if we assume that the integration and summation can be interchanged, then the preceding equation can be rewritten as

$$p_r(nT_b) = \int_{-1/2T_b}^{1/2T_b} \left(\sum_k P_r\left(f + \frac{k}{T_b}\right)\right) \exp(j2\pi fnT_b)\, df$$

Finally, if (5.4) is satisfied, then

$$\begin{aligned} p_r(nT_b) &= \int_{-1/2T_b}^{1/2T_b} T_b \exp(j2\pi fnT_b)\, df \\ &= \frac{\sin(n\pi)}{n\pi} \end{aligned}$$

which verifies that the $p_r(t)$ with a transform $P_r(f)$ satisfying (5.4) produces zero ISI.

The condition for the removal of ISI given in Equation (5.4) is called the Nyquist (pulse shaping) criterion.

Theorem 5.1 gives the condition for the removal of ISI using a $P_r(f)$ with a bandwidth larger than $r_b/2$. Proceeding along similar lines, it can be shown that ISI cannot be removed if the bandwidth of $P_r(f)$ is less than $r_b/2$.

The condition stated in Equation (5.4) does not uniquely specify $P_r(f)$. The particular choice of $P_r(f)$ for a given application is guided by two important considerations: the rate of decay of $p_r(t)$ and the ease with which shaping filters can be built. A pulse with a fast rate of decay and smaller values near $\pm T_b, \pm 2T_b, \ldots$ is desirable since these properties will yield a system in which modest timing errors will not cause large intersymbol interference. The shape of $P_r(f)$ determines the ease with which shaping filters can be realized. A $P_r(f)$ with a smooth roll-off characteristic is preferable over one with arbitrarily sharp cut-off characteristics, since the latter choice might lead to filters that will be hard to realize.

In practical systems where the bandwidth available for transmitting data at a rate of r_b bits/sec is between $r_b/2$ to r_b Hz, a class of $P_r(f)$ with a *raised cosine frequency characteristic* is most commonly used. A raised cosine frequency spectrum consists of a flat amplitude portion and a roll-off portion that has a sinusoidal form. The pulse spectrum $P_r(f)$ is specified in terms of a parameter β as

$$P_r(f) = \begin{cases} T_b, & |f| \leq r_b/2 - \beta \\ T_b \cos^2 \dfrac{\pi}{4\beta}\left(|f| - \dfrac{r_b}{2} + \beta\right), & \dfrac{r_b}{2} - \beta < |f| \leq \dfrac{r_b}{2} + \beta \\ 0, & |f| > r_b/2 + \beta \end{cases} \tag{5.6}$$

where $0 < \beta < r_b/2$. The pulse shape $p_r(t)$ corresponding to the $P_r(f)$ given above is

$$p_r(t) = \frac{\cos 2\pi\beta t}{1 - (4\beta t)^2} \left(\frac{\sin \pi r_b t}{\pi r_b t}\right) \tag{5.7}$$

Plots of $P_r(f)$ and $p_r(t)$ for three values of the parameter β are shown in

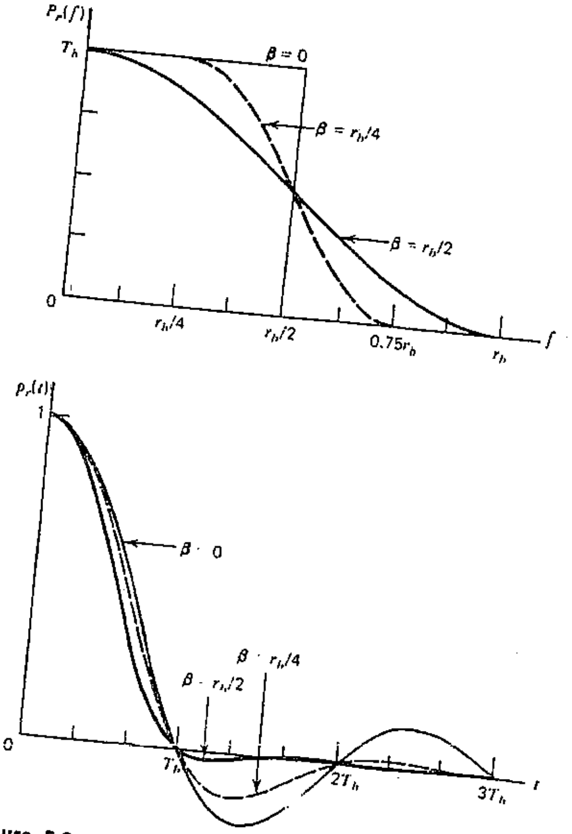


Figure 5.3 Pulses with raised cosine frequency characteristics. (a) $P_r(f)$ for three values of β. Note that $P_r(f) = P_r(-f)$. (b) $p_r(t)$ for three values of β

Figure 5.3. From Equations (5.6) and (5.7) and from Figure 5.3, the following observations can be made:

1. The bandwidth occupied by the pulse spectrum is $B = r_b/2 + \beta$. The minimum value of B is $r_b/2$ and the maximum value is r_b.
2. Larger values of β imply that more bandwidth is required for a given bit rate r_b. However, larger values of β lead to faster decaying pulses, which means that synchronization will be less critical and modest timing errors will not cause large amounts of ISI.

3. $\beta = r_b/2$ leads to a pulse shape with two convenient properties: the half amplitude pulse width is equal to T_b, and there are zero crossings at $t = \pm(\frac{3}{2})T_b, \pm(\frac{5}{2})T_b, \ldots$ in addition to zero crossings at $\pm T_b, \pm 2T_b, \ldots$. These properties aid in generating a timing signal for synchronization from the received signal.
4. $P_r(f)$ is real, nonnegative and $\int_{-\infty}^{\infty} P_r(f)\, df = 1$.

Summarizing the preceding discussion on the selection of a received pulse shape $p_r(t)$, we can say that a bandwidth of at least $r_b/2$ is required to generate a $p_r(t)$ producing zero ISI at a data rate r_b bits/sec. If additional bandwidth is available, then a $p_r(t)$ with an appropriate raised cosine spectrum given by (5.6) can be chosen. One must in general try to utilize all the available bandwidth up to r_b, and take advantage of faster decay of $p_r(t)$ with time. It must be pointed out that, strictly speaking, none of the raised cosine pulse spectra is physically realizable. A realizable frequency characteristic must have a time response that is zero prior to a time $t_0 (t_0 > 0)$, which is not the case for the $P_r(f)$ given in Equation (5.6). However, a delayed version of $p_r(t)$, say $p_r(t - t_d)$, may be generated by causal filters if the delay t_d is chosen such that $p_r(t - t_d) \approx 0$ for $t < t_0$. A practical filter that generates such a waveform is given in Problem 5.11.

5.2.2 Optimum Transmitting and Receiving Filters

The transmitting and receiving filters are chosen to provide proper pulse shaping and noise immunity. One of the design constraints that we have for selecting the filters is the relationship between the Fourier transforms of $p_r(t)$ and $p_g(t)$,

$$P_g(f)H_T(f)H_c(f)H_R(f) = K_c P_r(f) \exp(-j2\pi f t_d) \tag{5.8}$$

where t_d is the time delay* in the system and K_c is a normalizing constant. In order to design optimum filters $H_T(f)$ and $H_R(f)$, we will assume that $P_r(f)$, $H_c(f)$, and $P_g(f)$ are known.

If we choose $p_r(t)$ to produce zero ISI, then the constraint in (5.8) specifies that the filters shape $p_g(t)$ to yield a delayed version of $p_r(t)$. Now we need only to be concerned with noise immunity, that is, we need to choose the transmit and receive filters to minimize the effect of noise. For a given data rate, transmitter power, noise power spectral density, $H_c(f)$, and $P_r(f)$, we

*If t_d is sufficiently large, then the response of the system $K_c p_r(t - t_d)$ may be assumed to be 0 for $t < t_0$, where t_0 is the time at which the input $p_g(t)$ is applied to the system. Hence the filters will be causal.

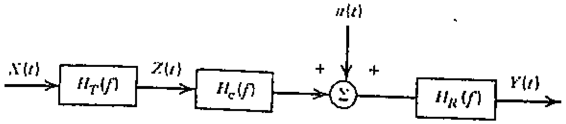


Figure 5.4 Portion of a baseband PAM system.

want to choose $H_T(f)$ and $H_R(f)$ such that the bit error probability is minimized (see Figure 5.4).

As a first step in deriving the optimum filters, let us derive an expression for the probability of a bit error. At the mth sampling time, the input to the A/D converter is

$$Y(t_m) = A_m + n_0(t_m), \quad t_m = mT_b + t_d$$

and the A/D converter output is 1 or 0 depending on whether $Y(t_m) > 0$ or < 0, respectively.* If we denote the mth input bit by d_m, then we can write an expression for the probability of incorrectly decoding the mth bit at the receiver as

$$\begin{aligned} P_e &= P[Y(t_m) > 0 | d_m = 0]P(d_m = 0) \\ &\quad + P[Y(t_m) < 0 | d_m = 1]P(d_m = 1) \end{aligned}$$

By virtue of our assumption of equiprobable bits, and due to the fact that $Y(t_m) = A + n_0(t_m)$ when $d_m = 1$, and $Y(t_m) = -A + n_0(t_m)$ when $d_m = 0$, where $A = K_c a$, we have

$$\begin{aligned} P_e &= \tfrac{1}{2}\{P[n_0(t_m) < -A] + P[n_0(t_m) > A]\} \\ &= \tfrac{1}{2}\{P[|n_0(t_m)| > A]\} \end{aligned}$$

The noise is assumed to be zero mean Gaussian at the input to $H_R(f)$, and hence the output noise $n_0(t)$ will also be zero mean Gaussian with a variance N_0 given by

$$N_0 = \int_{-\infty}^{\infty} G_n(f)|H_R(f)|^2\, df \tag{5.9}$$

*In general, for minimizing the probability of error, the receiver threshold should be set at

$$\frac{N_0}{2A}\log_e\frac{P(d_m = 0)}{P(d_m = 1)}$$

where $P(d_m = 0)$ and $P(d_m = 1)$ denote the probability that the mth input bit is 0 and 1, respectively, and N_0 is the variance of the noise at the input to the A/D converter (see Problem 5.5).

Using the above property, we can write P_e as

$$\begin{aligned} P_e &= \tfrac{1}{2}\int_{|x|>A} \frac{1}{\sqrt{2\pi N_0}} \exp(-x^2/2N_0)\, dx \\ &= \int_A^{\infty} \frac{1}{\sqrt{2\pi N_0}} \exp(-x^2/2N_0)\, dx \end{aligned} \tag{5.10}$$

A change of variable $z = x/\sqrt{N_0}$ yields

$$P_e = \int_{A/\sqrt{N_0}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)\, dz = Q\left(\frac{A}{\sqrt{N_0}}\right) \tag{5.11}$$

where

$$Q(u) = \int_u^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)\, dz$$

From Equation (5.11) we see that P_e decreases as $A/\sqrt{N_0}$ increases, and hence in order to minimize P_e we need to maximize the ratio $A/\sqrt{N_0}$. Thus, for maximum noise immunity the filter transfer functions $H_T(f)$ and $H_R(f)$ must be chosen to maximize the ratio $A/\sqrt{N_0}$. In order to do this maximizing, we need to express $A/\sqrt{N_0}$ or A^2/N_0 in terms of $H_T(f)$ and $H_R(f)$.

We start with the signal at the input to the transmitting filter

$$X(t) = \sum_{k=-\infty}^{\infty} a_k p_g(t - kT_b) \tag{5.12}$$

where $p_g(t)$ is a unit amplitude pulse having a duration less than or equal to T_b. Since the input bits are assumed to be independent and equiprobable, $X(t)$ is a random binary waveform with a psd (Chapter 3, Examples 3.8 and 3.9)

$$\begin{aligned} G_X(f) &= \frac{|P_g(f)|^2}{T_b} E\{a_k^2\} \\ &= \frac{a^2|P_g(f)|^2}{T_b} \end{aligned} \tag{5.13}$$

Now, the psd of the transmitted signal is given by

$$G_Z(f) = |H_T(f)|^2 G_X(f)$$

and the average transmitted power S_T is

$$S_T = \frac{a^2}{T_b}\int_{-\infty}^{\infty} |P_g(f)|^2 |H_T(f)|^2\, df \tag{5.14}$$

Since $A_k = K_c a_k$ and $A = K_c a$, we can write

$$S_T = \frac{A^2}{K_c^2 T_b}\int_{-\infty}^{\infty} |P_g(f)|^2 |H_T(f)|^2\, df$$

or

$$A^2 = K_c^2 S_T T_b \left[\int_{-\infty}^{\infty} |P_g(f)|^2 |H_T(f)|^2 \, df\right]^{-1} \tag{5.15}$$

Now, the average output noise power or the variance of $n_0(t)$ is given by

$$N_0 = \int_{-\infty}^{\infty} G_n(f) |H_R(f)|^2 \, df \tag{5.16}$$

and hence the quantity we need to maximize, A^2/N_0, can be expressed as

$$\frac{A^2}{N_0} = S_T T_b \left[\int_{-\infty}^{\infty} |H_R(f)|^2 G_n(f) \, df \int_{-\infty}^{\infty} \frac{|P_r(f)|^2}{|H_c(f) H_R(f)|^2} \, df\right]^{-1} \tag{5.17}$$

Or, we need to minimize

$$\gamma^2 = \int_{-\infty}^{\infty} |H_R(f)|^2 G_n(f) \, df \int_{-\infty}^{\infty} \frac{|P_r(f)|^2}{|H_c(f) H_R(f)|^2} \, df \tag{5.18}$$

The minimization of the right-hand side of (5.18) with respect to $H_R(f)$ can be carried out by using Schwarz's inequality, which is stated as follows: If $V(f)$ and $W(f)$ are complex functions of f, then

$$\int_{-\infty}^{\infty} |V(f)|^2 \, df \int_{-\infty}^{\infty} |W^*(f)|^2 \, df \geq \left|\int_{-\infty}^{\infty} V(f) W^*(f) \, df\right|^2 \tag{5.19}$$

The minimum value of the left-hand side of the equality is reached when $V(f) = \text{const times } W(f)$. Applying (5.19) to (5.18) with

$$|V(f)| = |H_R(f)| G_n^{1/2}(f)$$

$$|W(f)| = \frac{|P_r(f)|}{|H_c(f)| |H_R(f)|}$$

we see that γ^2 is minimized when

$$|H_R(f)|^2 = \frac{K |P_r(f)|}{|H_c(f)| G_n^{1/2}(f)} \tag{5.20}$$

where K is an arbitrary positive constant. Substituting Equation (5.20) in (5.8), we can obtain the optimum transmitting filter transfer function as

$$|H_T(f)|^2 = \frac{K_c^2 |P_r(f)| G_n^{1/2}(f)}{K |P_g(f)|^2 |H_c(f)|} \tag{5.21}$$

These filters should have linear phase response resulting in a total time delay of t_d (Equation (5.8)).

Finally, we obtain the maximum value of A^2/N_0 as

$$\left(\frac{A^2}{N_0}\right)_{\max} = (S_T)(T_b) \left[\int_{-\infty}^{\infty} \frac{|P_r(f)| G_n^{1/2}(f)}{|H_c(f)|} \, df\right]^{-2} \tag{5.22}$$

by substituting (5.20) in (5.17). The bit error probability P_e is then equal to

$$P_e = Q(\sqrt{(A^2/N_0)_{\max}}) \tag{5.23}$$

A special case of significant practical interest occurs when the channel noise is white $(G_n(f) = \eta/2)$, Gaussian, and when $P_g(f)$ is chosen such that it does not change much over the bandwidth of interest. The filter transfer functions now reduce to

$$|H_R(f)|^2 = K_1 \frac{|P_r(f)|}{|H_c(f)|} \tag{5.24a}$$

and

$$|H_T(f)|^2 = K_2 \frac{|P_r(f)|}{|H_c(f)|} \tag{5.24b}$$

where K_1 and K_2 are positive constants. From Equations (5.24a) and (5.24b) it follows that $|H_T(f)| = K_3 |H_R(f)|$, where K_3 is a positive constant. With the exception of an arbitrary gain difference, the transmitting and receiving filters have the same frequency characteristics so that one design serves both filters. In a large data communication system, having identical transmitting and receiving filters makes production and maintenance easy. A simple pulse shape $p_g(t)$ that yields an approximately constant $P_g(f)$ over the bandwidth of interest is

$$p_g(t) = \begin{cases} 1 & \text{for } |t| < \tau/2; \quad \tau \ll T_b \\ 0 & \text{elsewhere} \end{cases} \tag{5.25}$$

That is, a rectangular pulse of width $\tau \ll T_b$ can be used at the input to the transmit filter.

5.2.3 Design Procedure and Example

We will now see how the relationships derived in the preceding sections can be used to design a binary baseband PAM system given the bit rate r_b, acceptable error probability P_e, channel transfer function $H_c(f)$, and the channel noise power spectral density $G_n(f)$. Unless otherwise specified we will assume that the input bits are independent and equiprobable, that the channel noise is zero mean Gaussian, and that the channel is lowpass with a bandwidth B $(r_b/2 \leq B \leq r_b)$. If the channel bandwidth is less than $r_b/2$, we will have to use an M-ary signaling scheme that will be discussed later. If the channel bandwidth is much greater than r_b, then it would be wise to use some nonlinear modulation scheme to utilize the full bandwidth for reducing the effects of channel noise.

The design of the system consists of specifying the pulse shapes and spectra $P_r(f)$, $P_g(f)$, the transmitting and receiving filters $H_T(f)$, $H_R(f)$, and the transmitter power requirements S_T to meet the specified error probability. The steps involved in the design procedure are illustrated in the following example.

Example 5.1. Design a binary baseband PAM system to transmit data at a bit rate of 3600 bits/sec with a bit error probability less than 10^{-4}. The channel response is given by

$$H_c(f) = \begin{cases} 10^{-2} & \text{for } |f| < 2400 \\ 0 & \text{elsewhere} \end{cases}$$

The noise power spectral density is $G_n(f) = 10^{-14}$ watt/Hz.

Solution. We are given $r_b = 3600$ bits/sec, $P_e \leqslant 10^{-4}$, channel bandwidth $B = 2400$ Hz, and $G_n(f) = 10^{-14}$ watt/Hz.

If we choose a raised cosine pulse spectrum with $\beta = r_b/6 = 600$, then the channel bandwidth constraint is satisfied. Hence,

$$P_r(f) = \begin{cases} \dfrac{1}{3600}, & |f| < 1200 \\ \dfrac{1}{3600}\cos^2\dfrac{\pi}{2400}(|f| - 1200), & 1200 \leqslant |f| < 2400 \\ 0, & |f| \geqslant 2400 \end{cases}$$

Let us choose a $p_g(t)$ to satisfy (5.24) as

$$p_g(t) = \begin{cases} 1, & |t| < \tau/2 \\ 0, \quad \text{elsewhere:} & \tau = T_b/10 = (0.28)(10^{-4}) \end{cases}$$

Then,

$$P_g(f) = \tau\left(\frac{\sin \pi f\tau}{\pi f\tau}\right)$$

$$P_g(0) = \tau, \qquad P_g(2400) = 0.973\tau \approx \tau$$

Hence, the variation of $P_g(f)$ over $0 < |f| < 2400$ is very small and we obtain $|H_T(f)|$ and $|H_R(f)|$ from Equation (5.20) and (5.21) as

$$|H_T(f)| = K_1|P_r(f)|^{1/2}$$

$$|H_R(f)| = |P_r(f)|^{1/2}$$

We will choose $K_1 = (3600)(10^3)$ so that the overall response of H_T, H_c, and H_R to $P_g(f)$ produces $P_r(f)$, that is,

$$P_g(f)|H_T(f)||H_c(f)||H_R(f)| = P_r(f)$$

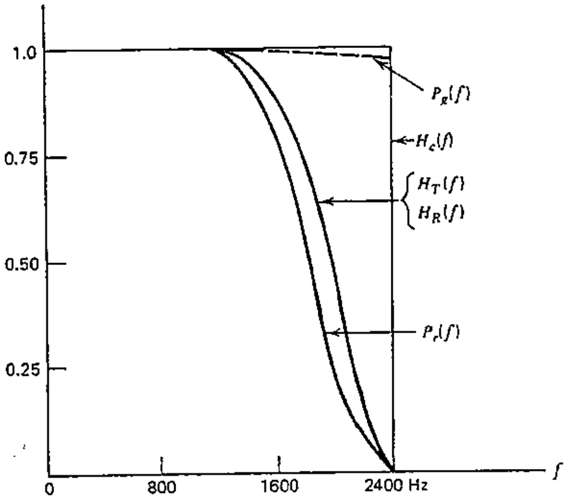


Figure 5.5 Normalized plots of P_g, H_c, H_T, H_R, and P_r. All functions are shown normalized with respect to their values at $f = 0$.

Plots of $P_g(f)$, $H_c(f)$, $H_T(f)$, $H_R(f)$, and $P_r(f)$ are shown in Figure 5.5.

Now, to maintain a $P_e \leqslant 10^{-4}$, we need $(A^2/N_0)_{\max}$ such that

$$Q(\sqrt{(A^2/N_0)_{\max}}) \leqslant 10^{-4}$$

Using the tabulated values of Q given in Appendix D we get

$$\sqrt{(A^2/N_0)_{\max}} \geqslant 3.75$$

or

$$(A^2/N_0)_{\max} \geqslant 14.06$$

From Equation (5.22) we obtain the transmitted power S_T as

$$\begin{aligned} S_T &= \frac{1}{T_b}\left(\frac{A^2}{N_0}\right)_{\max}\left[\int_{-\infty}^{\infty}\frac{|P_r(f)|G_n^{1/2}(f)}{|H_c(f)|}\,df\right]^2 \\ &= (3600)(14.06)\left(\frac{10^{-14}}{10^{-4}}\right)\left[\int_{-\infty}^{\infty}|P_r(f)|\,df\right]^2 \end{aligned}$$

For $P_r(f)$ with raised cosine shape $\int_{-\infty}^{\infty}|P_r(f)|\,df = 1$ and hence

$$S_T = (14.06)(3600)(10^{-10}) \approx -23\ \text{dBm}$$

which completes the design.

5.3 DUOBINARY BASEBAND PAM SYSTEM

In the preceding section we saw that a baseband binary PAM data transmission system requires a bandwidth of at least $r_b/2$ Hz in order to transmit data at a rate of r_b bits/sec, with zero ISI. If the bandwidth available is exactly $r_b/2$, then the only possible way in which binary PAM data transmission at a rate of r_b bits/sec can be accomplished without ISI would be to use ideal (rectangular) lowpass filters at the transmitter and receiver. Of course, such filters are physically unrealizable. Furthermore, any system that would closely approximate these filters would be extremely sensitive to perturbations in rate, timing, or channel characteristics.

In the past few years a class of signaling schemes known as *duobinary, polybinary,* or *partial response signaling* schemes has been developed to overcome some of the difficulties mentioned in the preceding paragraph. The duobinary scheme utilizes *controlled amounts of ISI* for transmitting data at a rate of r_b bits/sec over a channel with a bandwidth of $r_b/2$ Hz. The shaping filters for the duobinary system are easier to realize than the ideal rectangular filters needed to accomplish the data transmission at the maximum rate with zero ISI. The main disadvantage of the duobinary scheme is that it requires more power than an ideal binary PAM data transmission scheme.

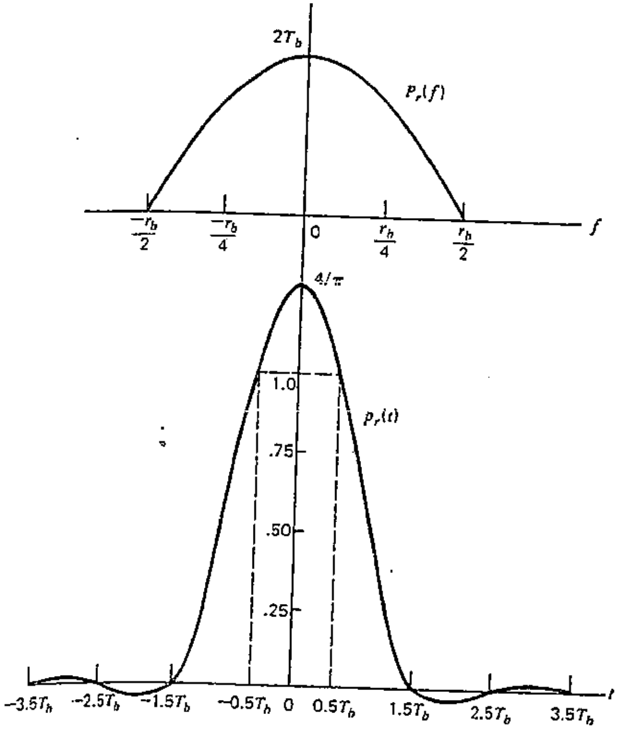


Figure 5.6 $P_r(f)$ and $p_r(t)$ for duobinary signaling scheme. Observe that the sampling is done at $t = (m \pm 0.5)T_b$.

The duobinary signaling schemes use pulse spectra $P_r(f)$ that yield $Y(t_m) = A_m + A_{m-1}$, where A_m and A_{m-1} are amplitudes related to the input bits d_m and d_{m-1}. One such $P_r(f)$ is

$$P_r(f) = \begin{cases} 2T_b \cos(\pi f T_b), & |f| \le 1/2T_b \\ 0, & |f| > 1/2T_b \end{cases} \qquad (5.26)$$

The pulse response $p_r(t)$ corresponding to the above $P_r(f)$ is

$$p_r(t) = \frac{4\cos(\pi t/T_b)}{\pi(1 - 4t^2/T_b^2)} \qquad (5.27)$$

Plots of $P_r(f)$ and $p_r(t)$ are shown in Figure 5.6. (See Problem 5.12 for other examples of $P_r(f)$.)

5.3.1 Use of Controlled ISI in Duobinary Signaling Scheme

The output $Y(t)$ of the receive filter can be written as

$$Y(t) = \sum_k A_k p_r(t - t_d - kT_b) + n_0(t) \qquad (5.28)$$

where $p_r(t)$ is defined in (5.27).

If the output is sampled at $t_m = mT_b - T_b/2 + t_d$, then it is obvious that in the absence of noise

$$Y(t_m) = A_m + A_{m-1} \qquad (5.29)$$

Equation (5.29) shows that the duobinary signaling scheme using a $P_r(f)$ given in Equation (5.26) introduces ISI. However, the intersymbol interference is controlled in that the interference comes only from the preceding symbol.

The A_m's in Equation (5.29) can assume one of two values, $\pm A$, depending on whether the mth input bit is 1 or 0. Since $Y(t_m)$ depends on A_m and A_{m-1}, $Y(t_m)$ can have one of the following *three* values (assuming no noise):

$$Y(t_m) = \begin{cases} +2A & \text{if the } m\text{th and } (m-1\text{st) bits are both 1's} \\ 0 & \text{if the } m\text{th and } (m-1\text{st) bits are different} \\ -2A & \text{if the } m\text{th and } (m-1\text{st) bits are both zero} \end{cases}$$

That is, the receiving filter output is a *three-level* waveform. The decoding of the mth bit from the sampled value of $Y(t_m)$ is done by checking to see if the input to the A/D converter at the sampling time is at the top, bottom, or middle level.

One apparent drawback of the system is that errors tend to propagate. Since the mth bit is decoded based on the decoded value of the $(m-1\text{st})$ bit, any error in the $(m-1\text{st})$ bit is likely to introduce an error in the decoding of the mth bit. A method of avoiding error propagation was proposed by Lender. In his scheme, error propagation is eliminated by *precoding* the input bit stream at the transmitter. The input bit stream (coming from the source) $b_1, b_2, b_3, \ldots$ is converted to another binary stream $d_1, d_2, d_3, \ldots$ before transmission according to the rule

$$d_m = b_m \oplus d_{m-1} \tag{5.30}$$

(The symbol $\oplus$ stands for modulo-2 addition.) The binary sequence d_k is transmitted using two levels $+a$ and $-a$ for 1 and 0, respectively.

If the mth input bit b_m is 0, then $d_m = d_{m-1}$ and according to Equation (5.29) $Y(t_m)$ will be $2A$ or $-2A$. On the other hand, if $b_m = 1$, then d_m will be the complement of d_{m-1} and $Y(t_m)$ will be zero. Hence the mth input bit b_m can be decoded according to the rule

$$\begin{aligned} b_m &= 0 \quad \text{if } Y(t_m) = \pm 2A \\ b_m &= 1 \quad \text{if } Y(t_m) = 0 \end{aligned} \tag{5.31}$$

In the preceding rule the mth bit is decoded from the value of $Y(t_m)$ only, and error propagation does not occur. Also, the implementation of the decoding algorithm is simple; $Y(t)$ is rectified and a simple threshold binary decision with a threshold at level A yields the output.

5.3.2 Transmitting and Receiving Filters for Optimum Performance

The procedure used for deriving optimum transmitting and receiving filters for the duobinary signaling scheme is the same as the one used in Section 5.2.2. The received levels at the input to the A/D converter are $2A$, 0, and $-2A$ with probabilities $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$, respectively. The probability of a bit error P_e is given by (assuming that the threshold is set at $\pm A$)

$$\begin{aligned} P_e &= \tfrac{1}{4}P\{n_0 < -A\} + \tfrac{1}{2}P\{|n_0| > A\} + \tfrac{1}{4}P\{n_0 > A\} \\ &= \tfrac{3}{2}P\{n_0 > A\} \end{aligned} \tag{5.32}$$

Since n_0 is a zero mean Gaussian random variable with a variance N_0', we can write P_e as

$$P_e = \tfrac{3}{2}Q(A/\sqrt{N_0'}) \tag{5.33}$$

The transmitting and receiving filters are chosen to maximize A^2/N_0' (in order to minimize P_e) subjected to the constraint that the overall pulse response $p_r(t)$ and its transform $P_r(f)$ satisfy Equations (5.26) and (5.27). It can be easily shown that the expressions for $H_T(f)$ and $H_R(f)$ will be the same as Equations (5.20) and (5.21).

The error probability P_e for the duobinary scheme will be higher than the error probability for the binary PAM system discussed in Section 5.1. For comparison purposes, let us assume that both schemes operate at the same bit rate r_b over an ideal channel $(H_c(f) = 1)$ with additive Gaussian white noise. For the direct binary PAM case it can be verified from Equation (5.22) that

$$\left(\frac{A^2}{N_0}\right)_{\max} = S_T T_b\left(\frac{2}{\eta}\right)$$

where $\eta/2 = G_n(f)$ is the noise power spectral density. Hence, the probability of error is

$$(P_e)_{\text{binary}} = Q\left(\sqrt{2\frac{S_T T_b}{\eta}}\right) \tag{5.34}$$

For the duobinary case from Equation (5.22), we have

$$\left(\frac{A^2}{N_0'}\right)_{\max} = S_T T_b \frac{2}{\eta}\left[\int_{-\infty}^{\infty} |P_r(f)|\, df\right]^{-2}$$

where $P_r(f)$ is given in Equation (5.20). The integral in the preceding equation can be evaluated as

$$\begin{aligned} \int_{-\infty}^{\infty} |P_r(f)|\, df &= \int_{-1/2T_b}^{1/2T_b} (2T_b)\cos(\pi f T_b)\, df \\ &= \frac{2}{\pi}\int_{-\pi/2}^{\pi/2} \cos\theta\, d\theta = \frac{4}{\pi} \end{aligned}$$

Hence,

$$\left(\frac{A^2}{N_0'}\right)_{\max} = 2\left(\frac{S_T T_b}{\eta}\right)\left(\frac{\pi}{4}\right)^2$$

and the probability of bit error for the duobinary scheme is given by Equation (5.33) as

$$(P_e)_{\substack{\text{duobinary}\\\text{PAM}}} = \tfrac{3}{2}Q\left(\frac{\pi}{4}\sqrt{2\frac{S_T T_b}{\eta}}\right) \tag{5.35}$$

A comparison of Equations (5.34) and (5.35) reveals that the bit error probability for the duobinary scheme is always higher than the bit error probability for an ideal binary PAM scheme using pulses with raised cosine frequency characteristics. However, the duobinary scheme uses less bandwidth than binary PAM schemes.

Example 5.2. Compare a binary PAM system with the duobinary system for transmitting data at a rate of 4000 bits/sec over an ideal channel ($H_c(f) = 1$) with Gaussian white noise. Assume $G_n(f) = \eta/2 = 10^{-12}$ watt/Hz and an error probability of 10^{-3}

Solution. For comparison let us consider binary PAM using a pulse with raised cosine frequency characteristics and $\beta = r_b/2$.

The bandwidth used is $r_b = 4000$ Hz.

$$P_e = Q\left(\sqrt{2\frac{S_T T_b}{\eta}}\right) < 10^{-3}$$

or

$$\sqrt{\frac{2S_T T_b}{\eta}} > 3.1$$

$$S_T > (3.1)^2\left(\frac{\eta}{2}\right)\frac{1}{T_b} \approx -44.2\,\text{dBm}$$

For the duobinary case, the bandwidth used is $r_b/2 = 2000$ Hz.

$$P_e = \tfrac{3}{2}Q\left(\frac{\pi}{4}\sqrt{\frac{2S_T T_b}{\eta}}\right) < 10^{-3}$$

or

$$\sqrt{\frac{2S_T T_b}{\eta}} > \left(\frac{4}{\pi}\right)(3.25) \quad \text{or} \quad S_T = -41.7\,\text{dBm}$$

The duobinary scheme uses $\frac{1}{2}$ the bandwidth of the binary PAM, but it requires about 2.5 dB more power than the binary PAM.

The discussion in the preceding section and the above example illustrate that the duobinary scheme can be used to transmit binary data at a rate of r_b bits/sec over a channel with a bandwidth of $r_b/2$ Hz. The shaping filters required by the scheme are easier to realize, and the pulse shape is such that small timing errors do not affect the performance of the system. The only comparable binary PAM system that can transmit r_b bits/sec over a channel with a bandwidth of $r_b/2$ Hz must use infinite cut-off ideal filters that are not physically realizable. Even if an approximation to the ideal filters can be realized, it would be extremely sensitive to any perturbations in timing or channel characteristics.

5.4 M-ARY SIGNALING SCHEMES

The baseband binary PAM systems discussed in the preceding sections use binary pulses, that is, pulses with one of two possible amplitude levels. In

M-ary baseband PAM systems, the output of the pulse generator is allowed to take on one of M possible levels ($M > 2$). Each level corresponds to a distinct input symbol, and there are M distinct input symbols. If symbols in the input sequence are equiprobable and statistically independent, then the information rate coming out of the pulse generator (Figure 5.7) is $r_s \log_2 M$ bits/sec, where r_s is the symbol rate.* Each pulse contains $\log_2 M$ bits of information. In the absence of noise and ISI, the receiver decodes input symbols correctly by observing the sampled values of the received pulse. As in the binary signaling scheme, noise and ISI introduce errors, and the major objectives of the design are then to eliminate ISI and reduce the effects of noise.

Before we discuss the design and analysis of M-ary signaling schemes, let us look at an example of such a scheme. Figure 5.7*a* shows the functional block diagram of an M-ary signaling scheme and Figure 5.7*b* shows the waveforms at various points in a *quarternary* ($M = 4$) system.

A comparison of Figures 5.7*a* and 5.7*b* with Figures 5.1 and 5.2 reveals that the binary and M-ary schemes are very similar, with the exception of the number of symbols in the input sequence and the corresponding number of amplitude levels of the pulses. In the quarternary scheme shown in Figure 5.7b, the source emits a sequence of symbols at a rate of r_s symbols/sec from a source alphabet consisting of four symbols A, B, C, and D. During each signaling interval of duration T_s ($= 1/r_s$), one of the four symbols is emitted by the source and the amplitude of the pulse generator output takes on one of four distinct levels. Thus, the sequence of symbols emitted by the source is converted to a four-level PAM pulse train by the pulse generator. The pulse train is shaped and transmitted over the channel, which corrupts the signal waveform with noise and distortion. The signal plus noise passes through the receiving filter and is sampled by the A/D converter at an appropriate rate and phase. The sampled value is compared against preset threshold values (also called *slicing levels*) and a decision is reached as to which symbol was transmitted, based on the sampled value of the received signal. Intersymbol interference, noise, and poor synchronization cause errors and the transmitting and receiving filters are designed to minimize the errors.

Procedures used for the design and analysis of multi-level PAM systems are similar to the ones we used for binary systems. We will assume that the input to the system is a sequence of statistically independent, equiprobable symbols produced by an ergodic source and that the channel noise is a zero-mean Gaussian random process with a psd of $G_n(f)$. We will see that an M-ary PAM scheme operating with the preceding constraints can transmit data at a bit rate of $r_s \log_2 M$ bits/sec and require a minimum bandwidth of $r_s/2$ Hz. In comparison, a binary scheme transmitting data at the same rate,

*The signaling speed is often given in the units of bauds.

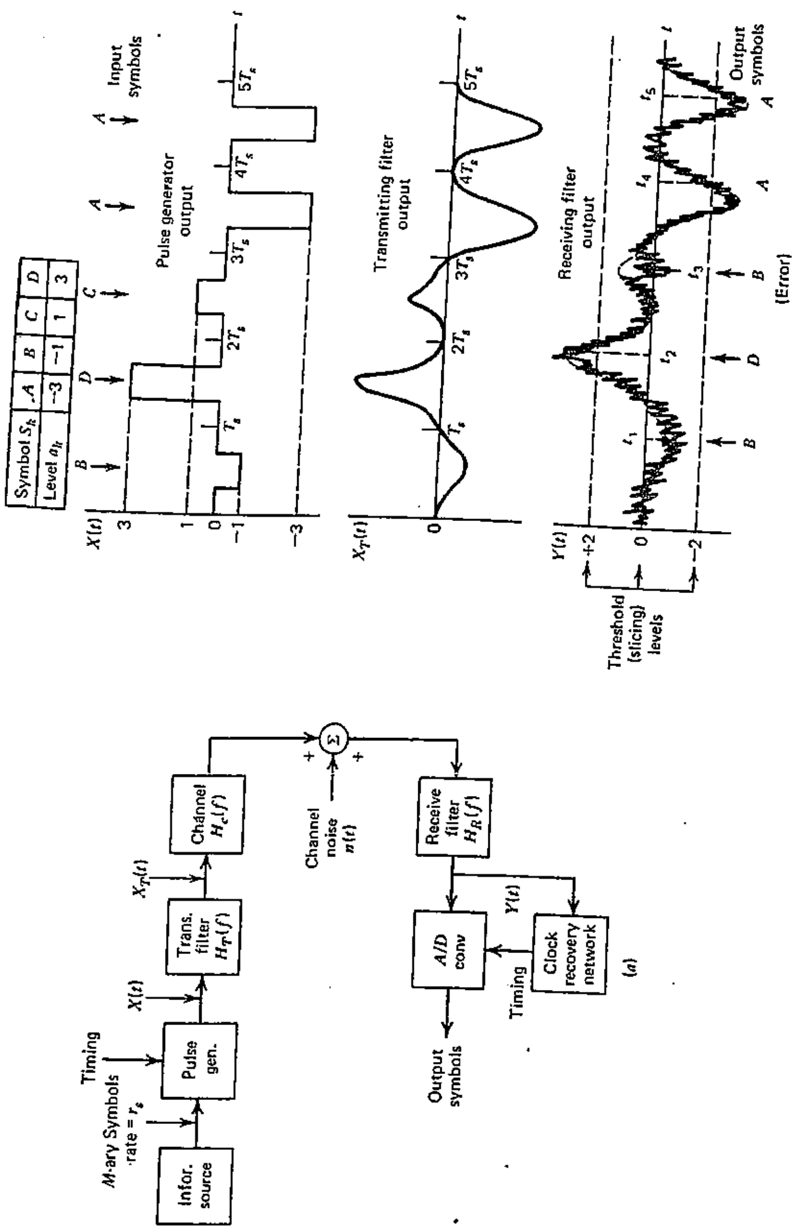


Figure 5.7 (a) Block diagram of an M-ary signaling scheme. (b) Signaling waveforms in a quarternary scheme.

$r_s \log_2 M$, will require a bandwidth of $(R_s \log_2 M)/2$ Hz. We will show that the price we pay for the bandwidth reduction that results from the use of M-ary schemes is the requirement for more power and more complex equipment.

5.4.1 Analysis and Design of *M*-ary Signaling Schemes

We start with the output of the pulse generator $X(t)$, which is given by

$$X(t) = \sum_{k=-\infty}^{\infty} a_k p_g(t - kT_s)$$

where $p_g(t)$ is the basic pulse whose amplitude a_k depends on the kth input symbol. If the source alphabet contains M letters, then a_k will take on one of M levels. We will assume that the spacing of amplitude levels is to be uniform and that a_k can take positive and negative values. It is easy to verify that for minimum power requirements, that is, for a minimum value of $E\{a_k - E\{a_k\}\}^2$, the M amplitude levels have to be centered at zero. For a separation of $2a$ between adjacent levels, the levels are given by

$$a_k = \begin{cases} 0, \pm 2a, \pm 4a, \ldots, \pm(M-1)a, & M \text{ odd} \\ \pm a, \pm 3a, \pm 5a, \ldots, \pm(M-1)a, & M \text{ even} \end{cases} \tag{5.36}$$

The signal power for the level spacings given in Equation (5.36) can be obtained from

$$E\{a_k^2\} = \frac{1}{M}\sum_{k=1}^{M}[(2k - M - 1)a]^2$$

When M is even, we have

$$E\{a_k^2\} = \frac{2}{M}\{a^2 + (3a)^2 + (5a)^2 + \cdots + [(M-1)a]^2\}$$

$$= \frac{(M^2 - 1)a^2}{3} \tag{5.37}$$

A similar expression can be derived for the case when M is odd.

Following the procedure used for the binary case (Section 5.2) we can write the input to the A/D converter at the sampling time as (from Equations (5.1) and (5.2))

$$Y(t_m) = A_m + \sum_{k \neq m} A_k p_r[(m-k)T_s] + n_0(t_m) \tag{5.38}$$

where $t_m = mT_s + t_d$ and t_d is the total time delay in the system. A_m takes on one of M values given by

$$A_m = \begin{cases} 0, \pm 2A, \pm 4A, \ldots, \pm(M-1)A, & M \text{ odd, } A = K_c a \\ \pm A, \pm 3A, \pm 5A, \ldots, \pm(M-1)A, & M \text{ even} \end{cases} \tag{5.39}$$

The second term in Equation (5.38) represents ISI, which can be eliminated by choosing a $p_r(t)$ with zero crossings at $\pm T_s, \pm 2T_s, \ldots$. We can use pulses with raised cosine frequency characteristics with the choice governed by rate of decay and bandwidth requirements. Pulses with raised cosine frequency characteristics have spectra $P_r(f)$ and waveform shapes $p_r(t)$ given by (Equation (5.6) with T_b replaced by T_s)

$$P_r(f) = \begin{cases} T_s, & |f| \leqslant r_s/2 - \beta \\ T_s \cos^2 \dfrac{\pi}{4\beta}\left(|f| - \dfrac{r_s}{2} + \beta\right), & \dfrac{r_s}{2} - \beta < |f| \leqslant \dfrac{r_s}{2} + \beta \\ 0, & |f| > r_s/2 + \beta \end{cases} \tag{5.40}$$

where $0 < \beta < r_s/2$, and

$$p_r(t) = \frac{\cos 2\pi\beta t}{1 - (4\beta t)^2}\left(\frac{\sin \pi r_s t}{\pi r_s t}\right) \tag{5.41}$$

with

$$\int_{-\infty}^{\infty} |P_r(f)|\, df = 1 \tag{5.42}$$

Since $P_r(f) = 0$ for $|f| > r_s/2 + \beta$, the channel bandwidth requirements for an M-ary signaling scheme is $r_s/2 + \beta$. With $0 \leqslant \beta \leqslant r_s/2$, the maximum BW requirement for an M-ary scheme using pulses with raised cosine frequency characteristics is r_s Hz. The data rate of such a scheme is $r_s \log_2 M$ bits/sec. For a binary scheme with the same data rate, the bandwidth requirement will be $r_s \log_2 M$.

The transmitting and receiving filters are chosen to produce zero ISI and minimize the probability of error for a given transmitted power. The zero ISI condition is met if $P_r(f)$ has the form given in Equation (5.40), and

$$P_g(f)H_T(f)H_c(f)H_R(f) = K_c \exp(-j2\pi f t_d)P_r(f) \tag{5.43}$$

where t_d is the total time delay, $P_g(f)$ is the transform of the basic pulse output of the pulse generator, and K_c is a normalizing constant. To minimize the probability of error, let us begin with the following expression for the probability density function of $Y(t_m)$. For convenience, let us use Y to denote $Y(t_m)$, and let us use the case where $M = 4$, shown in Figure 5.8. The probability density function of Y is given by

$$\begin{aligned} f_Y(y) = P(A_m = -3A)f_{Y|A_m=-3A}(y) + P(A_m = -A)f_{Y|A_m=-A}(y) \\ + P(A_m = A)f_{Y|A_m=A}(y) + P(A_m = 3A)f_{Y|A_m=3A}(y) \end{aligned} \tag{5.44}$$

where $f_{Y|A_m=kA}(y)$ is the conditional pdf of Y given that $A_m = kA$.

From Equation (5.38), with zero ISI,

$$Y_m = A_m + n_0$$

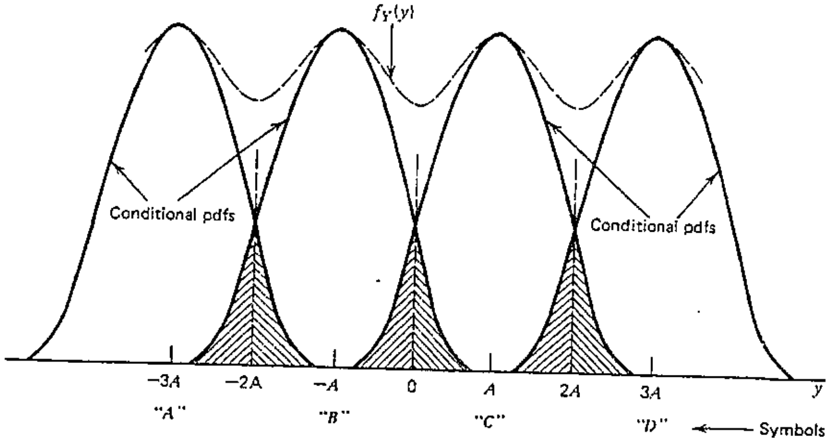


Figure 5.8 Probability density function of *signal pulse noise* for M = 4.

where n_0 is the noise term that has a zero mean Gaussian pdf

$$f_{n_0}(z) = \frac{1}{\sqrt{2\pi N_0}} \exp(-z^2/2N_0), \quad -\infty < z < \infty \tag{5.45}$$

and

$$N_0 = \int_{-\infty}^{\infty} G_n(f)|H_R(f)|^2\, df$$

Hence, for a given value of A_m, say $A_m = 3A$,

$$Y = 3A + n_0$$

and

$$f_{Y|A_m=3A}(y) = \frac{1}{\sqrt{2\pi N_0}} \exp(-(y - 3A)^2/2N_0), \quad -\infty < y < \infty$$

The conditional probability density functions corresponding to the four input symbols (hence, the four levels $-3A, -A, A, 3A$) are shown in Figure 5.8. By virtue of our assumption that the symbols are equiprobable, the optimum threshold levels are midway between the values of A_m, that is, the optimum decoding algorithm at the receiver is given by

$$\begin{array}{lll} \text{if} & Y(t_m) > 2A & \text{output symbol} = D \\ & 0 < Y(t_m) \leqslant 2A & \text{output symbol} = C \\ & -2A < Y(t_m) \leqslant 0 & \text{output symbol} = B \\ & Y(t_m) \leqslant -2A & \text{output symbol} = A \end{array}$$

The probability of error can now be written as

$$\begin{aligned}P(\text{error}) = {} & P(\text{error}|D \text{ was sent})P(D \text{ sent})\\ & + P(\text{error}|C \text{ was sent})P(C \text{ sent})\\ & + P(\text{error}|B \text{ was sent})P(B \text{ sent})\\ & + P(\text{error}|A \text{ was sent})P(A \text{ sent})\end{aligned}$$

or

$$\begin{aligned}P_e = {} & P(Y(t_m) \leq 2A|A_m = 3A)P(D \text{ sent})\\ & + P(Y(t_m) > 2A \quad \text{or} \quad \leq 0|A_m = A)P(C \text{ sent})\\ & + P(Y(t_m) > 0 \quad \text{or} \quad \leq -2A|A_m = -A)P(B \text{ sent})\\ & + P(Y(t_m) > -2A|A_m = -3A)P(A \text{ sent})\end{aligned} \tag{5.46}$$

Let us look at the second term in Equation (5.46) in detail. We have

$$P(C \text{ sent}) = \tfrac{1}{4}$$

and

$$\begin{aligned}& P(Y(t_m) > 2A \quad \text{or} \quad \leq 0|A_m = A)\\ & \quad = P(Y(t_m) > 2A|A_m = A) + P(Y(t_m) \leq 0|A_m = A)\\ & \quad = \int_{y>2A} f_{Y|A_m=A}(y)\, dy + \int_{y\leq 0} f_{Y|A_m=A}(y)\, dy\\ & \quad = \int_{y\geq 2A} \frac{1}{\sqrt{2\pi N_0}} \exp(-(z-A)^2/2N_0)\, dz + \int_{y\leq 0} \frac{1}{\sqrt{2\pi N_0}} \exp(-(z-A)^2/2N_0)\, dz\end{aligned}$$

A change of variable $u = (z-A)/\sqrt{N_0}$ yields

$$\begin{aligned}P(Y(t_m) > 2A \text{ or } \leq 0|A_m = A) &= 2\int_{u>A/\sqrt{N_0}} \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)\, du\\ &= 2Q\left(\frac{A}{\sqrt{N_0}}\right)\end{aligned}$$

Using a similar expression for the remaining terms in Equation (5.46), we arrive at an expression for P_e as

$$P_e = \tfrac{1}{4}(6)Q(A/\sqrt{N_0})$$

In the preceding expression, the factor $\frac{1}{4}$ represents the probability of occurrence of each of the four symbols; the factor 6 represents the six areas shown marked in Figure 5.8, and $Q(A/\sqrt{N_0})$ represents the numerical value of each one of the areas. Extending the derivation to the M-ary case and using similar arguments, we can obtain an expression for the probability of error as

$$P_e = \frac{2(M-1)}{M} Q\left(\frac{A}{\sqrt{N_0}}\right) \tag{5.47}$$

A comparison of Equation (5.47) with the probability of error for the binary case (given in Equation (5.11)) reveals that the two expressions are identical with the exception of the factor $2(M-1)/M$. The probability of error in the M-ary case is minimized when A^2/N_0 is maximized and the maximizing is carried out using a procedure similar to the one for the binary case (described in Section 5.2). Some of the intermediate results in the derivation are given below:

1. Transmitted power $= S_T$

$$= \frac{A^2}{K_c^2 T_s}\left(\frac{M^2-1}{3}\right)\int_{-\infty}^{\infty} |H_T(f)P_g(f)|^2\, df \tag{5.48}$$

2.

$$\frac{A^2}{N_0} = \left(\frac{3S_T T_s}{M^2-1}\right)\left[\int_{-\infty}^{\infty} |H_R(f)|^2 G_n(f)\, df \int_{-\infty}^{\infty} \frac{|P_r(f)|^2}{|H_c(f)H_R(f)|^2}\, df\right]^{-1} \tag{5.49}$$

3. The optimum transmitting and receiving filters that maximize A^2/N_0 given above are

$$|H_R(f)|^2 = \frac{K|P_r(f)|}{|H_c(f)|G_n^{1/2}(f)} \tag{5.50a}$$

$$|H_T(f)|^2 = \frac{K_c^2|P_r(f)|G_n^{1/2}(f)}{K|P_g(f)|^2|H_c(f)|} \tag{5.50b}$$

where K is a positive constant. The filter phase responses are arbitrary as long as Equation (5.43) is satisfied.

4.

$$\left(\frac{A^2}{N_0}\right)_{\max} = \frac{3S_T T_s}{M^2-1}\left[\int_{-\infty}^{\infty} \frac{|P_r(f)|G_n^{1/2}(f)}{|H_c(f)|}\, df\right]^{-2} \tag{5.51}$$

and

$$P_e = 2\left(\frac{M-1}{M}\right) Q\left(\sqrt{\left(\frac{A^2}{N_0}\right)_{\max}}\right) \tag{5.52}$$

5. In the special case when

$$G_n(f) = \eta/2$$

and $p_g(t)$ is chosen as

$$p_g(t) = \begin{cases} 1, & |t| < \tau/2;\ (\tau \ll T_s)\\ 0, & \text{elsewhere}\end{cases}$$

the filter transfer functions reduce to

$$|H_R(f)|^2 = K_1 \frac{|P_r(f)|}{|H_c(f)|} \tag{5.53}$$

$$|H_T(f)|^2 = K_2 \frac{|P_r(f)|}{|H_c(f)|} \tag{5.54}$$

where K_1 and K_2 are positive constants. As in the binary case, we have identical frequency response characteristics for the transmitting and receiving filters. Now if we also have

$$|H_c(f)| = \begin{cases} 1, & |f| < r_s/2 + \beta \\ 0, & \text{elsewhere} \end{cases}$$

(i.e., an ideal lowpass channel) then,

$$|H_R(f)| = |H_T(f)| = K_3|P_r(f)|^{1/2}$$

where K_3 is a positive constant. In this case we also have

$$\left(\frac{A^2}{N_0}\right)_{\max} = \left(\frac{3S_T T_s}{M^2-1}\right)\left(\frac{2}{\eta}\right) \tag{5.55}$$

and

$$P_e = \left(\frac{2(M-1)}{M}\right) Q\left(\sqrt{\left(\frac{3S_T T_s}{M^2-1}\right)\left(\frac{2}{\eta}\right)}\right) \tag{5.56}$$

The design and analysis of M-ary systems are carried out using equations (5.48)–(5.56).

Example 5.3. Design a quarternary signaling scheme for transmitting the output of an ergodic source emitting an independent sequence of symbols from a source alphabet consisting of four equiprobable symbols A, B, C, and D. The symbol rate of the source is 5000 symbols per sec and the overall error probability has to be less than 10^{-4}. The channel available for transmitting the data has the following characteristics:

$$H_c(f) = \frac{1}{1 + j(f/5000)}$$

and

$$G_n(f) = 10^{-12} \text{ watt/Hz}$$

Solution. We are given $M = 4$, $r_s = 5000$, and $T_s = (2)(10^{-4})$. Also we are given

$$|H_c(f)| = [1 + (f/5000)^2]^{-1/2}$$

Since there are no strict channel bandwidth limitations, we can use a received pulse having raised cosine frequency spectrum with $\beta = r_s/2$:

$$P_r(f) = \begin{cases} T_s \cos^2(\pi f/2r_s), & |f| < r_s \\ 0, & \text{elsewhere} \end{cases}$$

In order to arrive at $H_T(f)$ and $H_R(f)$ with similar functional forms, we choose $p_g(t)$ as

$$p_g(t) = \begin{cases} 1, & |t| < \tau/2 \\ 0, & \text{elsewhere} \end{cases} \quad \tau = T_s/10 = (2)(10^{-5})$$

$P_g(f)$ for the above $p_g(t)$ can be computed as

$$P_g(f) \approx \tau \quad \text{for } |f| < r_s$$

From Equations (5.50a) and (5.50b), we obtain $H_T(f)$ and $H_R(f)$ as

$$|H_R(f)| = \begin{cases} K_1\,[1 + (f/5000)^2]^{1/4} \cos(\pi f/2r_s), & |f| < 5000 \text{ Hz} \\ 0, & \text{elsewhere} \end{cases}$$

$$|H_T(f)| = \begin{cases} [1 + (f/5000)^2]^{1/4} \cos(\pi f/2r_s), & |f| < 5000 \text{ Hz} \\ 0, & \text{elsewhere} \end{cases}$$

The constant K_1 and the phase shifts of the filters are chosen to yield

$$P_g(f)H_T(f)H_c(f)H_R(f) = P_r(f)\exp(-j2\pi f t_d)$$

Plots of $P_g(f)$, $H_c(f)$, $H_T(f)$, $H_R(f)$, and $P_r(f)$ are shown in Figure 5.9.

Now we need to find the transmitter power requirements to maintain $P_e < 10^{-4}$. From Equation (5.52), we have

$$P_e = 2 \cdot \tfrac{3}{4} Q(\sqrt{(A^2/N_0)_{\max}}) < 10^{-4}$$

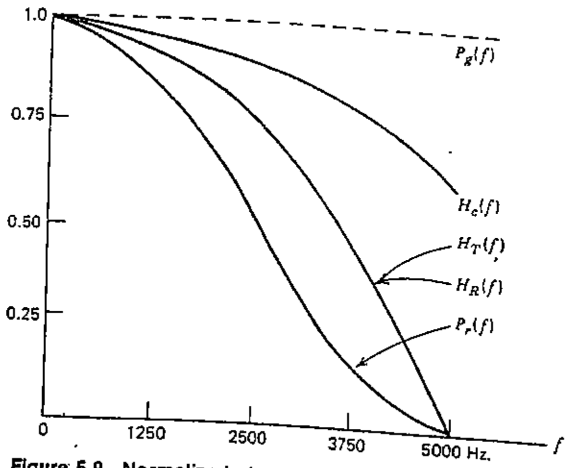


Figure 5.9 Normalized plots of P_g, H_c, H_T, H_R, and $P_r(f)$. All functions are shown normalized with respect to their values at $f = 0$.

which requires

$$Q(\sqrt{(A^2/N_0)_{\max}}) < (\tfrac{2}{3})(10^{-4})$$

or

$$(A^2/N_0)_{\max} > (3.8)^2 = 14.44$$

From Equation (5.51), we obtain

$$\left(\frac{A^2}{N_0}\right)_{\max} = \frac{3S_T T_s}{15}\left[\int_{-\infty}^{\infty} \frac{|P_r(f)|G_n^{1/2}(f)}{|H_c(f)|}\,df\right]^{-2} > 14.44$$

The integral in the preceding equation can be evaluated either in closed form (when possible) or by a numerical approximation:

$$\begin{aligned}\int_{-\infty}^{\infty} \frac{|P_r(f)|G_n^{1/2}(f)}{|H_c(f)|}\,df &= 10^{-6}\int_{-5000}^{5000} T_s\left(\cos\frac{\pi f}{2r_s}\right)^2\left[1+\left(\frac{f}{5000}\right)^2\right]^{1/2} df\\ &= (2)(10^{-6})\int_0^1 \left(\cos\frac{\pi}{2}x\right)^2 (1+x^2)^{1/2}\,dx\\ &\approx (1.20)10^{-6} \quad \text{(by numerical integration)}\end{aligned}$$

Hence

$$\left(\frac{S_T T_s}{5}\right)[(1.20)10^{-6}]^{-2} > 14.44$$

or

$$\begin{aligned}S_T &> (5)(5000)(1.20)^2(10^{-12})(14.44)\\ &> (-32.5)\ \text{dBm}\end{aligned}$$

5.4.2 Binary versus *M*-ary Signaling Schemes

We are now ready to compare binary and M-ary signaling schemes and determine which scheme should be used in a particular situation. Let us assume that:

1. The input to both systems comes from an ergodic information source that emits an independent sequence of equally likely binary digits at a rate of r_b bits/sec (no loss in generality results from this assumption since the output symbol sequence of any source may be coded into a binary sequence).
2. The channel is ideal lowpass, and the channel noise is zero mean Gaussian with a psd of $\eta/2$.
3. Both signaling schemes use pulses having appropriate raised cosine frequency characteristics with the maximum value of β.
4. Both signaling schemes are required to yield the same error probability P_e.

Table 5.1. Comparison of binary and *M*-ary signaling schemes

	Binary scheme	M-ary scheme
Bandwidth	r_b Hz	$r_s = (r_b/k)$ Hz
Probability of Error P_e	$Q\left(\sqrt{\frac{S_T}{r_s \log_2 M}\left(\frac{2}{\eta}\right)}\right)$ (Eqs. (5.22); (5.23))	$\frac{2(M-1)}{M}Q\left(\sqrt{\frac{3S_T}{(M^2-1)r_s}\left(\frac{2}{\eta}\right)}\right)$ (Eq. (5.56))
Transmitter power for a given P_e	Less	More
Equipment complexity	Less	More

Bit rate = r_b bits/sec.
M-ary symbol rate = $r_s = r_b/k$ symbols/sec.

The M-ary signaling scheme is assumed to operate on blocks of k binary digits at a time ($M = 2^k$). Each block of k binary digits is translated to one of M levels at the transmitter, and each received level is decoded as a block of k binary digits at the receiver. The bandwidth and power requirements of the binary and M-ary schemes are shown in Table 5.1.

Comparison of binary and M-ary signaling schemes indicate that *binary transmission has lower power requirements*, and that *M-ary signaling schemes require lesser bandwidth*. For $M \gg 2$ and $P_e \ll 1$, the transmitter power must increase by a factor of $M^2/\log_2 M$, whereas the bandwidth is reduced by $1/\log_2 M$. M-ary schemes are more complex since the receiver has to decide on one of M-levels using $M-1$ comparators or level slicers. In the binary case, the decoding requires only one comparator.

Example 5.4. Compare the power-bandwidth requirements of binary and quarternary ($M = 4$) signaling schemes for transmitting the output of an ergodic source emitting an independent sequence of symbols from an alphabet consisting of four equiprobable letters A, B, C, and D. The symbol rate is 5000/sec and the signaling schemes are required to maintain $P_e \le 10^{-4}$. Assume an ideal lowpass channel with additive Gaussian noise with a psd $\eta/2 = 10^{-12}$ watt/Hz.

Solution. The data rate for the problem is $5000 \log_2 4 = 10{,}000$ bits/sec. To design a binary signaling scheme, we need to convert the symbols into bit strings. This can be done by assigning 2-bit binary words to each letter in the source alphabet. For example, we can assign 00, 01, 10, and 11 to A, B, C, and

D, respectively. A symbol sequence such as *DBCAADCB* will be translated into a bit stream 1101100000111001 and transmitted as binary data. The receiver will first decode the individual binary digits, and then decode the letters by looking at two bit groups.

Hence, for the binary scheme we have

$$r_b = 10{,}000 \text{ bits/sec}$$

$$H_c(f) = 1, \quad |f| < 10{,}000 \text{ Hz}$$

$$G_n(f) = \eta/2 = 10^{-12}$$

If we use a received pulse having raised cosine frequency characteristics with $\beta = r_b/2$, then the bandwidth required is r_b Hz, that is,

$$(\text{Bandwidth})_{\text{binary}} = 10{,}000 \text{ Hz}$$

Power requirement may be computed using

$$P_e = Q(\sqrt{(A^2/N_0)_{\max}}) < 10^{-4}$$

which requires

$$(A^2/N_0)_{\max} \geqslant (3.75)^2 = 14.06$$

From Equation (5.22) with $G_n(f) = 10^{-12}$, $H_c(f) = 1$, and $\int_{-\infty}^{\infty} P_r(f)\,df = 1$, we get

$$\left(\frac{A^2}{N_0}\right)_{\max} = \frac{S_T T_b}{(\eta/2)} = S_T(10^8)$$

Hence

$$(S_T)_{\text{binary}} \geqslant (14.06)(10^{-8}) = -38.52 \text{ dBm}$$

For the quarternary ($M = 4$) scheme the bandwidth required is (using $P_r(f)$ with $\beta = r_s/2$)

$$(\text{Bandwidth})_{\text{quarternary}} = 5000 \text{ Hz}$$

and

$$P_e = \tfrac{3}{2}Q(\sqrt{(A^2/N_0)_{\max}}) < 10^{-4}$$

or

$$(A^2/N_0)_{\max} \geqslant 14.44$$

From Equation (5.51), we have

$$\left(\frac{A^2}{N_0}\right)_{\max} = \left(\frac{S_T T_s}{5}\right)\left(\frac{2}{\eta}\right)$$

$$(S_T)_{\text{quarternary}} \geqslant (14.44)(5\eta/2T_s)$$
$$\geqslant -34.42 \text{ dBm}$$

The reader can, for comparison purposes, verify that a scheme with $M = 16$ can be designed for this problem. The requirements for $M = 16$ are

$$(\text{Bandwidth})_{M=16} = 2500 \text{ Hz}$$

$$(S_T)_{M=16} \geqslant -24.9 \text{ dBm}$$

M	Bandwidth (Hz)	Power (dBm)
2	10,000	−38.52
4	5000	−34.42
16	2500	−24.90

The results in the table show that as the bandwidth is reduced the power requirement increases sharply.

In the preceding comparison of binary and M-ary PAM systems, the symbol error probability was used as the basis for comparison. If the output of the information source is assumed to be binary digits, it is more meaningful to use the bit error probability for comparing M-ary and binary schemes. While there are no unique relationships between the bit error probability and M-ary symbol error probability, we can derive such relationships for two special cases.

In the first case we assume that whenever an M-ary symbol is in error, the receiver output is equally likely to be any one of the $2^k - 1$ erroneous k-bit sequences. Now, considering any arbitrary bit position in the M-fold set of k-bit sequences, $M/2$ sequences contain a binary 1 in that position and the remaining sequences contain a binary 0. Then considering an arbitrary bit in the input sequence, it is apparent that the same binary digit occurs in $2^{k-1} - 1$ of the remaining $2^k - 1$ possible sequences and the other binary digit occurs in that position in the remaining 2^{k-1} sequences. Under these assumptions, the average probability of bit error P_{eb} in M-ary transmission with a symbol error probability P_e is given by

$$P_{eb} = \frac{2^{k-1}}{2^k - 1} P_e \leqslant \tfrac{1}{2}P_e$$

Thus in comparing binary and M-ary schemes, we should use

$$P_e = \begin{cases} P_{eb} & \text{for binary transmission} \\ P_{eb}\left(\dfrac{2^k - 1}{2^{k-1}}\right) & \text{for } M\text{-ary transmission} \end{cases} \tag{5.57a}$$

where P_{eb} is the bit error probability of both schemes.

In practical M-ary PAM schemes, one often uses a binary-to-M-ary coding such that binary sequences corresponding to adjacent amplitude levels differ in only one bit position (one such coding is the so-called Gray code). Now, when an M-ary symbol error occurs, it is more likely that only one of the k input bits will be in error, especially when the signal-to-noise ratio at the receiver input is high. In such a case, we have

$$P_{eb} = P_e/k$$

Thus, for a given bit error probability, we should use

$$P_e = \begin{cases} P_{eb} & \text{for binary transmission} \\ kP_{eb} & \text{for } M\text{-ary transmission} \end{cases} \tag{5.57b}$$

for comparison purposes.

5.5 SHAPING THE TRANSMITTED SIGNAL SPECTRUM

In many applications, the spectrum of the PAM signal should be carefully shaped to match the channel characteristics in order to arrive at physically realizable transmitting and receiving filters. As an example, consider a channel that has a poor amplitude response (high attenuation) in the low-frequency range. Using a signaling scheme that has high power content in the low-frequency range will result in transmitting and receiving filters with unreasonably high gain in the low-frequency range (see Equations (5.20) and (5.21)). This problem could be avoided if the spectrum of the transmitted signal is altered so that it has small power content at low frequencies.

The spectrum of the transmitted signal in a PAM system will depend on the signaling pulse waveform and on the statistical properties of sequences of transmitted digits. The pulse waveform in a PAM system is specified by the ISI requirements. While it might be possible to shape the transmitted signal spectrum by changing the transmitted pulse shape, such changes might lead to increased ISI. An easier way to shape the transmitted signal spectrum is to alter the statistical properties of the transmitted bit or symbol sequence.

In the following section we will look at methods of changing the spectrum of the transmitted signal by changing the statistical properties of the amplitude sequence of the pulse train. We will also look at digital methods of generating accurately shaped signaling waveforms.

5.5.1 Effect of Precoding on the Signal Spectrum

The output of the pulse generator in a PAM system is a pulse train $X(t)$, which can be written as

$$X(t) = \sum_{k=-\infty}^{\infty} a_k p_g(t - kT)$$

where T is the symbol duration ($T = T_b$ for binary schemes and $T = T_s$ for M-ary schemes). The output of the transmitting filter $Z(t)$ is also a pulse train,

$$Z(t) = \sum_{k=-\infty}^{\infty} a_k p_t(t - kT)$$

where $p_t(t)$ is the response of the transmitting filter to $p_g(t)$. The power spectral density $G_Z(f)$ of $Z(t)$ is given by (Chapter 3, Examples 3.8 and 3.9)

$$G_Z(f) = \frac{|P_t(f)|^2}{T} G(f) \tag{5.58a}$$

where $G(f)$ has the form

$$G(f) = R(0) - m^2 + 2 \sum_{k=1}^{\infty} [R(k) - m^2] \cos 2\pi k f T \tag{5.58b}$$

In the preceding equations $P_t(f)$ is the Fourier transform of $p_t(t)$, $R(k)$ represents $E\{a_j a_{j+k}\}$, and $m = E\{a_k\}$. The factor $G(f)$ represents the way in which the spectrum is affected by the statistics of the amplitude level sequence $\{a_k\}$. We will consider three commonly used methods of precoding which alter the statistical properties of the transmitted sequence and hence the spectrum of the transmitted signal. We will illustrate the principles of precoding using a binary PAM system with an input bit stream $\{b_k\}$ that is assumed to be an equiprobable, independent sequence from an ergodic source. The bit sequence $\{b_k\}$ is converted into an amplitude sequence $\{a_k\}$ which amplitude modulates $p_t(t - kT)$ (see Figure 5.10). We will look at methods of mapping $\{b_k\}$ into $\{a_k\}$ and their effects on the shape of the transmitted spectrum. We will concentrate our attention on $G(f)$ in Equation (5.58).

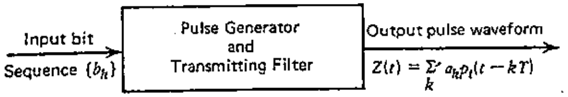


Figure 5.10 Portion of a binary PAM system.

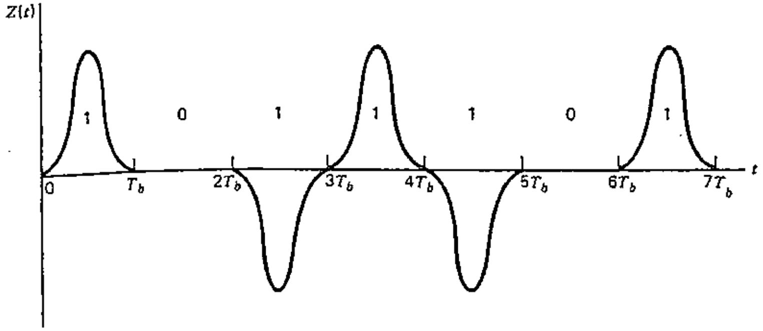


Figure 5.11 An example of bipolar coding.

Bipolar coding. The most widely used method of coding for baseband PAM data transmission is the *bipolar* coding. The bipolar coding uses three amplitude levels: zero and equal magnitude positive and negative levels (for convenience we will take these levels to be +1, 0, and −1). Binary 0 in the input is represented by $a_k = 0$, and binary 1 is represented by either +1 or −1. Each successive 1 that occurs is coded with opposite polarity as shown in Figure 5.11. The autocovariance $R(j)$ for this sequence can be calculated as follows:

$$R(0) = E\{a_k^2\}$$
$$= (1)^2P(a_k = 1) + (-1)^2P(a_k = -1) + (0)^2P(a_k = 0)$$

It can be easily verified that $P(a_k = 1) = P(a_k = -1) = \frac{1}{4}$ and $P(a_k = 0) = \frac{1}{2}$. Hence

$$R(0) = \tfrac{1}{2}$$

$R(1)$ and $R(2)$ can be calculated using the joint probabilities given in Table 5.2. .

The reader can verify that

$$R(1) = E\{a_k a_{k+1}\}$$
$$= \sum_{i=-1}^{1}\sum_{j=-1}^{1} ijP(a_k = i,\ a_{k+1} = j)$$
$$= -\tfrac{1}{4}$$

and

$$R(2) = E\{a_k a_{k+2}\}$$
$$= \sum_{i=-1}^{1}\sum_{j=-1}^{1} ijP(a_k = i,\ a_{k+2} = j)$$
$$= -\tfrac{1}{8} + \tfrac{1}{8} = 0$$

Table 5.2. Joint probabilities for $a_k a_{k+j}$

d_k	d_{k+1}	d_{k+2}	a_k	a_{k+1}	a_{k+2}	Prob.
0	0	0	0	0	0	$\frac{1}{8}$
0	0	1	0	0	1	$\frac{1}{8}$
0	1	0	0	1	0	$\frac{1}{8}$
0	1	1	0	1	−1	$\frac{1}{8}$
1	0	0	1	0	0	$\frac{1}{8}$
1	0	1	1	0	−1	$\frac{1}{8}$
1	1	0	1	−1	0	$\frac{1}{8}$
1	1	1	1	−1	1	$\frac{1}{8}$

(Assume the last 1 was coded as −1.)

a_{k+1} \ a_k	−1	0	1
−1	0	0	$\frac{1}{4}$
0	0	$\frac{1}{4}$	$\frac{1}{4}$
1	0	$\frac{1}{4}$	0

a_{k+2} \ a_k	−1	0	1
−1	0	$\frac{1}{8}$	$\frac{1}{8}$
0	0	$\frac{1}{4}$	$\frac{1}{4}$
1	0	$\frac{1}{8}$	$\frac{1}{8}$

(Entries inside these tables are joint probabilities.)

Following a similar procedure it can be shown that

$$R(j) = 0 \quad \text{for } j > 2$$

Hence the spectral component $G(f)$ for the bipolar coding becomes

$$G(f) = \tfrac{1}{2}(1 - \cos 2\pi f T_b) \tag{5.59}$$

A plot of $G(f)$ shown in Figure 5.12 reveals that this scheme can be used for baseband signaling without DC or near DC components (which makes it ideal for channels which have poor low-frequency response). Also, the average power is somewhat less since one half of the pulses have zero amplitude. However, the missing pulses will make clock recovery difficult.

Twinned Binary Coding. Another commonly used method of coding fo binary PAM data transmission is the twinned binary coding which is also called *split phase* or *Manchester coding*. In this method, each bit is represented by two successive pulses of opposite polarity, the two binary values

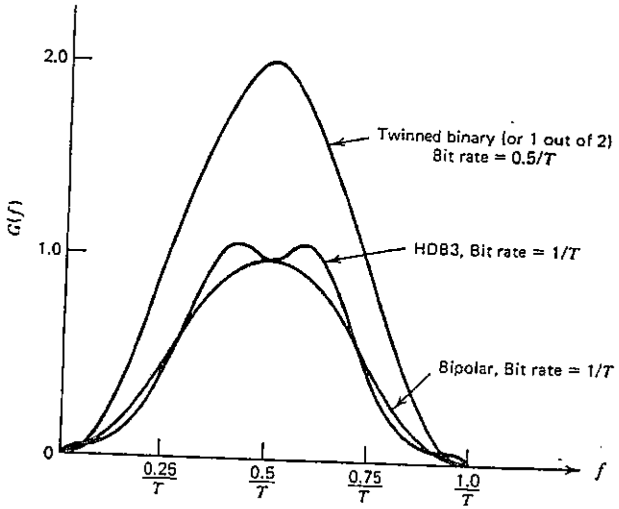


Figure 5.12 Power spectral density of the transmitted signal.

having the representations +− and −+ (Figure 5.13). The bit rate is half that of unrestricted binary pulses. Hence, this method uses twice the bandwidth. For the twinned binary scheme, $R(0) = 1$, $R(1) = -\frac{1}{2}$, $T = T_b/2$, and

$$G(f) = 1 - \cos \pi f T_b \tag{5.60}$$

A plot of $G(f)$ for the twinned binary coding is shown in Figure 5.12. Like the bipolar coding scheme, twinned binary coding also has no near DC components. The main advantage of the twinned binary coding scheme is that every pulse position is occupied and hence clock recovery is easier. The disadvantage is that the twinned binary scheme requires larger bandwidth.

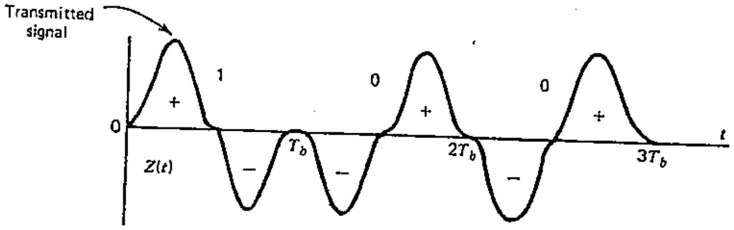


Figure 5.13 Twinned binary coding.

The twinned binary coding is also called a one out of two coding scheme since one out of every two pulses is positive and the other one is negative. An extension of this method leads to 2 out of 4, 3 out of 6, and 4 out of 8 schemes that produce spectra with sharper roll-offs. These methods are very useful for tailoring the transmitted signal spectrum to a finite frequency band.

High Density Bipolar Coding. A series of high density bipolar coding schemes known as HDB1, HDB2, HDB3, . . . are used to eliminate long strings of zero amplitude pulses that occur in regular bipolar coding schemes. Absence of signaling pulses make clock recovery difficult. In the regular bipolar coding scheme, successive pulses have opposite sign. In the HDB coding schemes, "bipolar" rule violations are used to carry extra information needed to replace strings of zeros. The HDB*n* scheme is designed to avoid the occurrence of more than *n* pulses of zero amplitude. The most important of the HDB codes is the HDB3.

The HDB3 code uses a bipolar coding scheme whenever possible. But if the string 0000 occurs in the input binary stream, special sequences other than 0000 are transmitted. The special sequences contain bipolar violations and hence can be easily distinguished. The special sequences used in HDB3 coding are 000D and 100D. In the special sequence, "1" is represented by amplitude level $a_k = +1$ or -1 *following the bipolar rule*, and "0" by level 0. The "D" is replaced by level +1 or −1 *violating the bipolar rule.* The choice of the special sequence 000D or 100D is made in such a way that the pulses violating the bipolar rule take on levels +1 and −1 alternately. 100D is used when there has been an even number of ones since the last special sequence. Special sequences can follow each other if the string of zeros continues. Two special sequences are necessary to assure that the special sequences can be distinguished from data sequences, and to guarantee that there will be zero crossings when special sequences follow each other. An example is shown below.

Input bit stream	1	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Coded bit stream	1	0	1	1	1	0	0	D	0	1	0	0	0	D	1	0	0	D
Amplitude levels a_k	−	0	+	−	+	0	0	+	0	−	0	0	0	−	+	0	0	+

(Special sequences are shown enclosed. The choice of the first special sequence is arbitrary.)

The HDB3 waveform corresponding to the above example is shown in Figure 5.14.

The calculation of the autocovariance and $G(f)$ for HDB codes is rather lengthy and hence is not included here. However, a plot of $G(f)$ for the HDB3

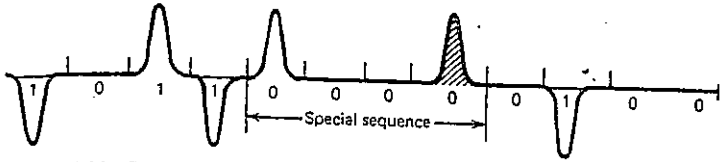


Figure 5.14 Example of HDB3 waveform. Shaded pulse represents a bipolar violation.

coding is shown in Figure 5.12 for comparison purposes. As may be expected, $G(f)$ for HDB3 lies close to the bipolar curve. The mean power for the HDB3 is 10% higher than the bipolar case since some of the 0 amplitude pulses in the bipolar case are replaced by nonzero amplitudes in HDB3 coding schemes.

5.5.2 Pulse Shaping by Digital Methods

Digital signaling schemes require accurately shaped pulse waveforms. The shape of a pulse can be specified either by its amplitude as a function of time or by its Fourier transform. If a pulse shape is specified by its transform in the frequency domain, then, in principle at least, we can design filters to shape the pulse. This design becomes difficult in a practical sense because of the need for linear phase characteristics for the filters. An alternate method is to generate the pulse waveform directly. Such a method should be capable of generating a signal composed of many overlapping pulses that are superposed. One of the most commonly used methods of direct waveform generation makes use of a binary transversal filter shown in Figure 5.15. The pulse waveform $p_t(t)$ which is to be amplitude modulated is sampled at intervals of Δ and the amplitude values at sampling times are $b_{-8}, b_{-7}, \ldots, b_{-1}, b_0, b_1, \ldots, b_8$. For purposes of simplicity we will assume that sample values outside the time interval -8Δ to 8Δ are small enough to be ignored. The transversal filter for this waveform consists of a 17-bit shift register with 17 outputs, each of which can be two possible levels 0 or 1. These outputs are attenuated by b_{-8} to b_8, respectively, and summed by an analog summing network.

The actual waveform we are trying to generate has the form

$$Z(t) = \sum_k d_k p_t(t - kT_b)$$

where d_k is the kth input bit having a value 0 or 1. In the example shown in Figure 5.14, we are approximating $Z(t)$ by the staircase waveform

$$\tilde{Z}(t) = \sum_k d_k \tilde{p}_t(t - kT_b)$$

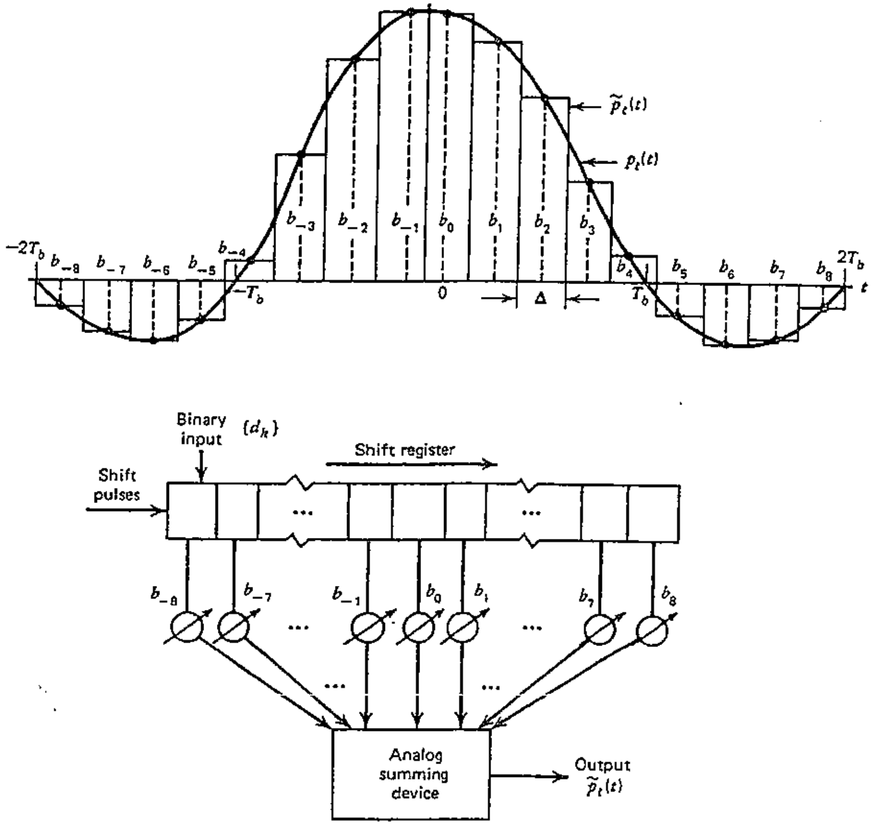


Figure 5.15 Waveform generation using a binary transversal filter.

that is, $p_t(t)$ is approximated by

$$\tilde{p}_t(t) = \sum_{j=-8}^{8} b_j p_s(t - j\Delta)$$

where

$$p_s(t) = \begin{cases} 1, & -\Delta/2 \leq t < \Delta/2 \\ 0, & \text{elsewhere} \end{cases}$$

The kth input bit d_k stays in the shift register for a duration of 17Δ seconds. It is shifted through the shift register at a rate equal to 17 times the bit rate.

Successive bits d_k are inserted into the shift register once every T_b seconds and the output due to successive input bits overlap and the summing device performs the superposition of individual pulses to form the composite waveform.

The digitally generated waveform has zero errors at sampling times 0, $\pm T_b$, $\pm 2T_b, \ldots$. The spectrum of the staircase waveform is centered around the sampling frequency ($1/\Delta$) and its harmonics. By choosing a suitably large sampling frequency the "noise" due to sampling can be separated from the spectrum of the desired pulses. Then, a simple lowpass filter can be used to smooth out the waveform.

For signaling methods with more than one amplitude level, we need additional shift registers attenuators, and summing devices. For example, if the signaling scheme demands three levels 1, 0, and -1, then two shift registers will be needed.

5.6 EQUALIZATION

In the design of baseband PAM systems we assumed that the frequency response $H_c(f)$ of the channel is completely known. Based on the knowledge of $H_c(f)$ we designed PAM systems to yield zero intersymbol interference (ISI). In almost all real systems some amount of residual ISI inevitably occurs due to imperfect filter design, incomplete knowledge of channel characteristics, changes in channel characteristics, and so forth. The only recourse to mitigate the residual distortion is to include within the system an adjustable filter or filters that can be "trimmed" to compensate for the distortion. The process of correcting channel induced distortion is called *equalization*. Equalizing filters are most often inserted between the receiving filter and the A/D converter, especially in systems using switched telephone lines where the specific line characteristics are not known in advance.

5.6.1 Transversal Equalizer

It is obvious that an equalizing filter should have a frequency response $H_{eq}(f)$ such that the actual channel response multiplied by $H_{eq}(f)$ yields the assumed channel response $H_c(f)$ that was used in the system design. However, since we are interested in the output waveform at a few predefined sampling times only, the design of an equalizing filter is greatly simplified. The most commonly used form of easily adjustable equalizer has been the transversal filter shown in Figure 5.16.

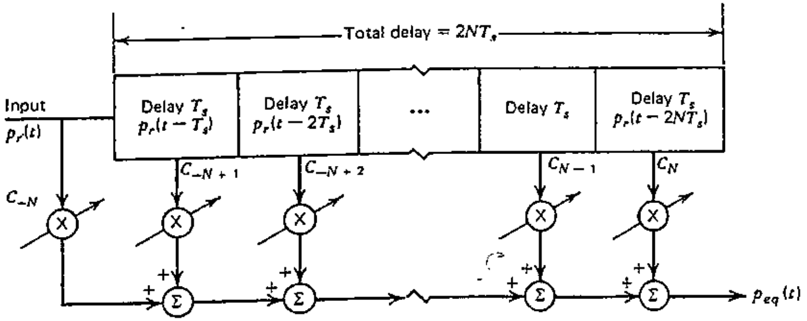


Figure 5.16 Transversal equalizer.

The transversal equalizer consists of a delay line tapped at T_s second intervals. Each tap is connected through a variable gain device to a summing amplifier. For convenience we will assume that the filter has $(2N+1)$ taps with gains C_{-N}, $C_{-N+1}, \ldots, C_0, C_1, \ldots, C_N$. The input to the equalizer is $p_r(t)$, which is known, and the output is $p_{eq}(t)$. We can write the output $p_{eq}(t)$ in terms of $p_r(t)$ and the tap gains as

$$p_{eq}(t) = \sum_{n=-N}^{N} C_n p_r[t-(n+N)T_s] \tag{5.61}$$

If $p_r(t)$ has its peak at $t=0$ and ISI on both sides, the output should be sampled at $t_k = (k+N)T_s$ and

$$p_{eq}(t_k) = \sum_{n=-N}^{N} C_n p_r[(k-n)T_s] \tag{5.62}$$

If we denote $p_r(nT_s)$ by $p_r(n)$ and t_k by k, we have

$$p_{eq}(k) = \sum_{n=-N}^{N} C_n p_r(k-n) \tag{5.63}$$

Ideally, we would like to have

$$p_{eq}(k) = \begin{cases} 1 & \text{for } k=0 \\ 0 & \text{elsewhere} \end{cases}$$

This condition cannot always be realized since we have only $(2N+1)$ variables (namely, the $2N+1$ tap gains) at our disposal. However, we can specify the value of $p_{eq}(t)$ at $2N+1$ points as

$$p_{eq}(k) = \begin{cases} 1 & \text{for } k=0 \\ 0, & k=\pm 1, \pm 2, \ldots, \pm N \end{cases} \tag{5.64}$$

Combining Equations (5.63) and (5.64), we have

$$\begin{array}{c} N \\ \text{zeros} \\ \\ \\ N \\ \text{zeros} \end{array} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{(2N+1)} = \begin{bmatrix} p_r(0) & p_r(-1) & \cdots & p_r(-2N) \\ p_r(1) & p_r(0) & \cdots & p_r(-2N+1) \\ & & & \vdots \\ p_r(2) & p_r(1) & \cdots & \vdots \\ \vdots & & & \vdots \\ p_r(2N) & \cdots & \cdots & p_r(0) \end{bmatrix} \begin{bmatrix} C_{-N} \\ C_{-N+1} \\ \vdots \\ C_0 \\ \cdot \\ C_{N-1} \\ C_N \end{bmatrix} \qquad (5.65)$$

Equation (5.65) represents a set of $(2N + 1)$ simultaneous equations that can be solved for the C_n's. The equalizer described in Equation (5.64) is called a zero forcing equalizer since $p_{eq}(k)$ has N zero values on either side. This equalizer is optimum in that it minimizes the peak intersymbol interference. The main disadvantage of a zero forcing equalizer is that it increases the noise power at the input to the A/D converter (Problem 5.21). But this effect is normally more than compensated for by the reduction in the ISI.

Example 5.5. Design a three tap equalizer to reduce the ISI due to the $p_r(t)$ shown in Figure 5.17*a*.

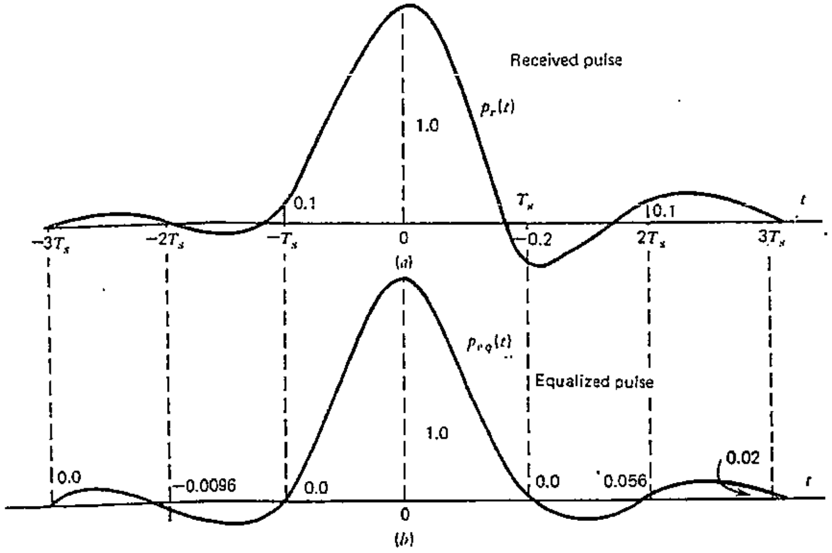


Figure 5.17 Three tap equalizer discussed in Example 5.5.

Solution. With a three tap equalizer we can produce one zero crossing on either side of $t = 0$ in the equalized pulse. The tap gains for this equalizer are given by

$$\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1.0 & 0.1 & 0 \\ -0.2 & 1.0 & 0.1 \\ 0.1 & -0.2 & 1.0 \end{pmatrix} \begin{pmatrix} C_{-1} \\ C_0 \\ C_1 \end{pmatrix}$$

or

$$\begin{pmatrix} C_{-1} \\ C_0 \\ C_1 \end{pmatrix} = \begin{pmatrix} -0.09606 \\ 0.9606 \\ 0.2017 \end{pmatrix}$$

The values of the equalized pulse can be computed from Equation (5.63) as

$$\begin{aligned} &p_{eq}(-3) = 0.0 && p_r(-3) = 0.0 \\ &p_{eq}(-2) = -0.0096 && p_r(-2) = 0.0 \\ &p_{eq}(-1) = 0.0 && p_r(-1) = 0.1 \\ &p_{eq}(0) = 1.0 && p_r(0) = 1.0 \\ &p_{eq}(1) = 0.0 && p_r(1) = -0.2 \\ &p_{eq}(2) = 0.0557 && p_r(2) = 0.1 \\ &p_{eq}(3) = 0.02016 && p_r(3) = 0.0 \end{aligned}$$

The equalized pulse is shown in Figure 5.17*b*. While the equalized pulse has one zero crossing on either side of $t = 0$, it has small ISI at points further out from $t = 0$ where the unequalized pulse had zero ISI.

5.6.2 Automatic Equalizers

The design and adjustment of the tap gains of the zero forcing equalizer described in the preceding section involves the solution of a set of simultaneous equations. In the manual mode, the "trimming" of the equalizer involves the following steps:

1. Send a test pulse through the system.
2. Measure the output of the receiving filter $p_r(t)$ at appropriate sampling times.
3. Solve for the tap gains using Equation (5.65).
4. Set the gains on the taps.

"Highly accurate and simply instrumented *automatic systems* for setting up the gains have been developed in recent years. These systems are usually divided into two groups: the *preset type* that uses a special sequence of pulses prior to or during breaks in actual data transmission, and the *adaptive type* that adjusts itself continuously during data transmission by operating on the data signal itself. Automatic equalizers use iterative techniques to arrive at optimum tap gains. Before we take a detailed look at the operation of automatic equalizers, let us briefly review an iterative technique commonly used for solving a set of simultaneous equations.

The equations we have to solve, given in (5.65), can be written as

$$I = XC \tag{5.66}$$

where I is a $(2N+1)$ column vector whose components are all zero except the zeroth, X is a $(2N+1)$ square matrix whose (i, j)th element is $p_r(i-j)$, and C is a column vector whose jth element C_j represents the gain of the jth tap; the indices i, j run from $-N$ to N. The iterative method assumes that at the end of the kth iteration we have a solution vector C^k that yields an error in the solution

$$\epsilon^k = XC^k - I \tag{5.67}$$

The components of the error vector are denoted by ϵ_j^k $(j=-N$ to $N)$. The new, adjusted value of the solution vector C^{k+1} is obtained by

$$C^{k+1} = C^k - \Delta\,\mathrm{sgn}(\epsilon^k) \tag{5.68}$$

where

$$\mathrm{sgn}(y) = \begin{cases} +1, & y > 0 \\ 0, & y = 0 \\ -1, & y < 0 \end{cases}$$

and Δ represents a positive increment. The iterations are continued until C^k and C^{k+1} differ by less than some arbitrarily small value.

Under some mild assumptions, it has been shown that the system distortion per tap can be reduced to near the minimum possible value of $\Delta/2$. In this method (called the fixed increment method) Δ must be made as small as possible. Two examples of automatic equalizers are presented next.

Preset Equalizer. A simple preset equalizer is shown in Figure 5.18. In this system, components of the error vector ϵ^k $(k=1, 2, \ldots)$ are measured by transmitting a sequence of widely separated pulses through the system and observing the output of the equalizer $(= XC^k)$ at the sampling times. Fixed increment iteration, with stepsize Δ, is used to adjust the tap gains. The sampling of the filter output is done using a timing circuit triggered by a peak detector. The center sample is sliced at $+1$ (or compared with $+1$) and the polarity of the error component ϵ_0^k is obtained. The polarities of the other error components ϵ_j^k $(j=-N$ to $N)$ are obtained from the value of the filter output at $t=\pm jT_s$. (If we denote the values of the equalized pulse at the end of the kth iteration by $p_{eq}^k(t)$, then $\epsilon_j^k=p_{eq}^k(jT_s)$ for $j\neq 0$, and $\epsilon_0^k=p_{eq}^k(0)-1$.)

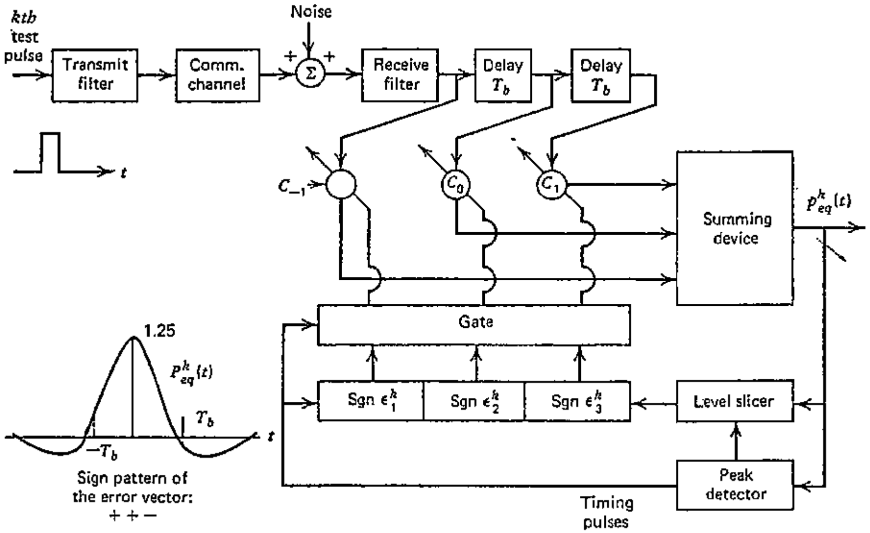


Figure 5.18 A three tap preset equalizer.

At the end of the kth test pulse, the gate is opened, and depending on the polarity of the components of ϵ^k, the tap gains are moved up or down by Δ according to (5.68). This iterative "training session" is carried on until the procedure converges. The training procedure might involve hundreds of pulses.

A major problem in "training" a preset equalizer is the presence of noise in the observed values of $p_{eq}(t)$. The effects of noise can be somewhat minimized by averaging the measured values of $p_{eq}(t)$ over a number of test pulses before the tap gains are changed. One of the difficulties with averaging is that the rate of convergence is slowed down somewhat. Other methods for reducing the effects of noise may be found in Reference 1 listed at the end of this chapter.

Adaptive Equalizer. In adaptive equalizers, the error vector ϵ^k is continually estimated during the normal course of data transmission. Such schemes have the ability to adapt to changes during data transmission and eliminate the need for long training sessions. Adaptive equalizers are quite

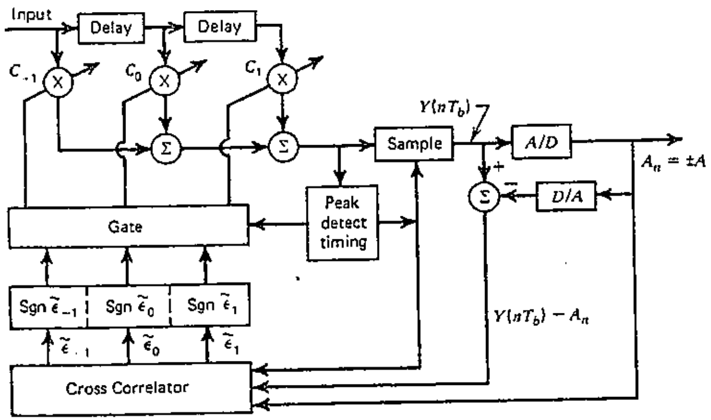


Figure 5.19 A three tap adaptive equalizer.

common in practice and are more accurate, versatile, and cheaper than preset equalizers.

A simple adaptive equalizer for a binary data transmission scheme is shown in Figure 5.19. The output of the equalizer $Y(t)$ at sampling times should be $\pm A$, $+A$ if the actual input bit corresponding to the sampling time is 1 and $-A$ if the input bit is 0. In an actual system, due to ISI, the values $Y(jT_b)$ will vary about $\pm A$ depending on the input sequence. For a random data sequence these variations will also be random. If the ISI is not very large, we can still decode the data and generate a sequence of ideal or desired levels A_j, where $A_j = \pm A$. From the ideal sequence A_j and the actual measured values $Y(jT_b)$, we can generate an estimate of the error sequence needed for adjusting the tap gains. The estimate* most commonly used is given by [Lucky, page 160]

$$\hat{\epsilon}_j = \frac{1}{A^2 m} \sum_{n=1}^{m} (A_{n-j})[Y(nT_b) - A_n] \tag{5.69}$$

where m is the sequence length used for estimation. The jth tap gain is adjusted according to $C_j^{k+1} = C_j^k - \Delta \operatorname{sgn}\{\hat{\epsilon}_j\}$; $j = -N$ to N, where k denotes the number of the iteration cycle.

In order for the adaptive equalizer to work effectively the input bit sequence and the sequence of received samples $Y(nT_b)$ must be random. Further, adaptive equalizers have a difficult time establishing initial equalization. Once correct equalization is acquired, the error estimates are accurate and the equalization loop tracks changes in channel characteristics easily. A procedure used often to circumvent this difficulty is to use a hybrid system in which data transmission is delayed during a brief reset period in which a quasi-random sequence is transmitted and regenerated at the receiver. When the equalization is reasonably good, the equalizer is switched to an adaptive mode and data transmission is started.

*$\hat{\epsilon}_j$ is the maximum likelihood estimator of ϵ_j, the jth component of the error vector. The right-hand side of Equation (5.69) represents the cross correlation operation.

5.7 MISCELLANEOUS TOPICS

5.7.1 Eye Diagram

The performance of baseband PAM systems depends on the amount of ISI and channel noise. The distribution of ISI and channel noise in the system can be readily observed by displaying the received waveform $Y(t)$ on an oscilloscope using a sweep rate that is a fraction of the symbol rate r_s. The resulting display shape resembles a human eye and is widely known as the *eye pattern* of the system. To understand and interpret eye patterns let us consider a binary PAM system. The received waveform with no noise and no distortion is shown in Figure 5.20*a*. When segments of this waveform are superimposed on each other, the "open" eye pattern results. A vertical line drawn through the center of the eye pattern reveals that if the sampling time is correct, then all sampled values are $\pm A$.

Figure 5.20*b* shows a distorted version of the waveform and the corresponding eye pattern. The eye pattern appears "closed" and the sampled values are now distributed about $\pm A$. Decoding of the received waveform is somewhat difficult now. Finally, 5.20*c* shows a noisy distorted version of the received waveform and the corresponding eye pattern. Plots shown in these figures reveal that the eye pattern displays useful information about the performance of the system. For comparison purposes typical eye patterns of a duobinary signaling scheme are shown in Figure 5.21.

Eye patterns are often used for monitoring the performance of baseband PAM systems. If the signal-to-noise ratio at the receiver is high, then the following observations can be made from the eye pattern shown simplified in Figure 5.22:

1. The best time to sample the received waveform is when the eye opening is largest.
2. The maximum distortion is indicated by the vertical width of the two branches at sampling time.

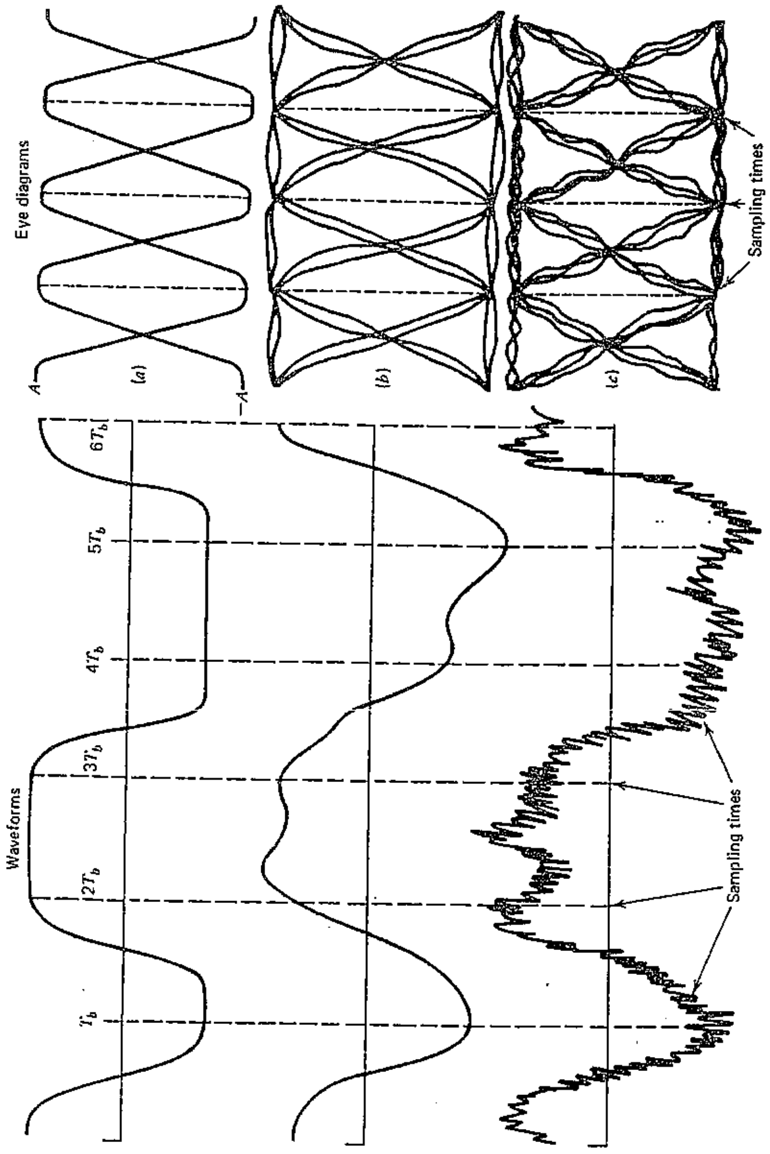
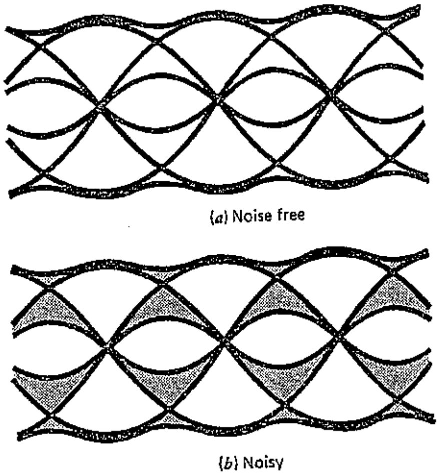


Figure 5.20 Eye diagrams of a binary PAM system. (a) Ideal. (b) Distorted. (c) Distortion + noise.

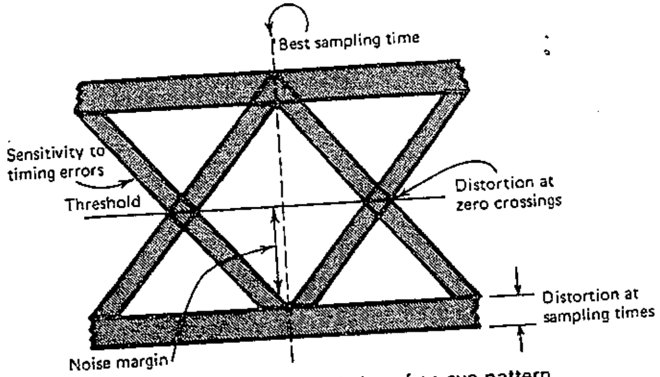


(a) Noise free

(b) Noisy

Figure 5.21 Eye pattern of a duobinary system.

3. The noise margin or immunity to noise is proportional to the width of the eye opening.
4. The sensitivity of the system to timing errors is revealed by the rate of closing of the eye as sampling time is varied.
5. The sampling time is midway between zero crossings; if the clock information is derived from zero crossings, then the amount of distortion of zero crossings indicates the amount of "*jitter*" or variations in clock rate and phase.



Best sampling time
Sensitivity to timing errors
Threshold
Distortion at zero crossings
Distortion at sampling times
Noise margin

Figure 5.22 Characteristics of an eye pattern.

6. Asymmetries in the eye pattern indicate nonlinearities in the channel since in a strictly linear system with truly random data all the eye openings will be identical.

5.7.2 Synchronization

In baseband PAM systems the output of the receiving filter $Y(t)$ must be sampled at precise sampling instants $t_m = mT_s + t_d$. To do this sampling we need a clock signal at the receiver that is synchronized with the clock signal at the transmitter. Three general methods in which this synchronization can be obtained are:

1. derivation of clock information from a primary or secondary standard; for example, the transmitter and receiver can be slaved to a master clock;
2. transmitting a synchronizing clock signal;
3. derivation of the clock signal from the received waveform itself.

The first method is often employed in large data communication networks. On point-to-point data transmission at low rates this method is not necessary, and the high cost of this method does not justify its use. The second method involves the transmission of a clock signal along with data which means that a small part of the channel's information capacity needs to be given over to the clock signal. If the available capacity is large compared to the data rate requirements, then this method is most reliable and least expensive.

The third method, *self-synchronization*, is a very efficient method of synchronizing the receiver to the transmitter. An example of a system used to derive a clock signal from the received waveform is shown in Figure 5.23.

The clock recovery network consists of a voltage controlled oscillator (VCO) and a phase comparator consisting of the phase comparison logic and transistor controlled current switches. The phase comparison logic circuit is triggered by the one shot multivibrator that outputs a pulse of duration $T_b/2$ when the input $Y(t)$ is ≤ 0. The correction or error signal comes out of the phase comparator in the form of I_c. The charging and discharging of the capacitor is controlled by I_c and the voltage across the capacitor controls the VCO, which generates the clock signal.

To illustrate the operation of the phase comparator network, let us look at the timing diagram shown in Figure 5.23*b*. At time t_1, the clock signal is in phase and between time t_1 and t_2 the clock signal drifts by a small amount. As $Y(t)$ goes negative at t_1, the one shot is triggered and it puts out a pulse of duration $0.5T_b$. The phase comparison logic generates two equal width pulses QC and $Q\bar{C}$, and the current I_c has a waveform with equal width, equal amplitude positive and negative portions. The net change in the charge across the capacitor is zero, the VCO control voltage remains constant, and no adjustment is made on the rate and phase of the clock signal.

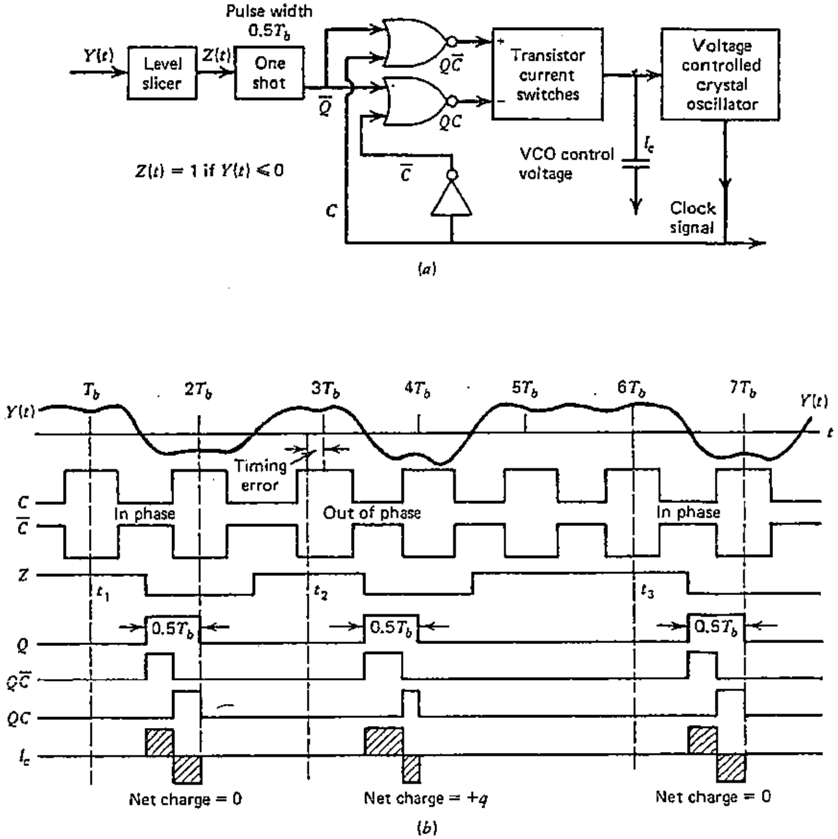


Figure 5.23 Clock recovery network.

As the clock signal drifts out of phase, the phase comparison operation at time t_2 results in a current pulse I_c with a more positive component. Now, there is a change q in the capacitor charge and hence a change in the VCO control voltage. This change in the VCO control voltage results in a correction of the clock phase. In the example shown in Figure 5.23*b*, the clock phase is shown corrected before the next phase comparison operation is initiated at t_3.

Several versions of the clock recovery network shown in Figure 5.23 are used in practice. Almost all self-synchronizing networks depend on level changes or zero crossings in the received waveform $Y(t)$. The performance of these networks will degrade considerably if the signal stays at a constant level for long periods of time and if the zero crossings in $Y(t)$ are obliterated by noise. The lack of level changes in the data can be corrected by using one of the coding methods suggested in the preceding section. The effect of noise on the zero crossing can be minimized by setting an error threshold in the phase comparator output below which no clock phase correction is attempted. Another way of minimizing the effect of noise is to use an averaged estimate of the zero crossing times in the phase comparison network.

5.7.3 Scrambler and Unscrambler

Binary communication systems are designed to convey a sequence of bits from the source to the receiver. While the system is expected to convey all possible sequences, there may be some sequences that are not conveyed correctly. For example, the clock recovery in the system might be affected if a long series of 1's is sent and hence the system might start losing data due to timing errors. A binary communication system is said to have *bit sequence transparency* if it can convey *any* given sequence of bits. Several methods are used to preserve bit transparency and most of the methods involve an encoding procedure to restrict the occurrence of periodic sequences and sequences containing long strings of ones or zeros.

Scrambler. Many subsystems in data communication systems work best with random bit sequences. Examples of such subsystems are adaptive equalizers and self-synchronization networks. While strings of ones or zeros, or periodic sequences might appear in the output of an information source, such sequences have to be recoded for transmission if the data transmission system has difficulty in conveying these sequences. A device commonly used for recoding undesirable bit strings is called a scrambler. While it may not be possible to prevent the occurrence of all undesirable sequences with absolute certainty, at least most of the common repetitions in the input data can be removed by the use of a scrambler.

The scrambler shown in Figure 5.24*a* consists of a "feedback" shift register and the matching unscrambler has a "feed forward" shift register structure. In both the scrambler and unscrambler the outputs of several stages of shift register are added together, modulo-2, and then added to the data stream again in modulo-2 arithmetic. The contents of the registers are shifted at the bit rate of the system.

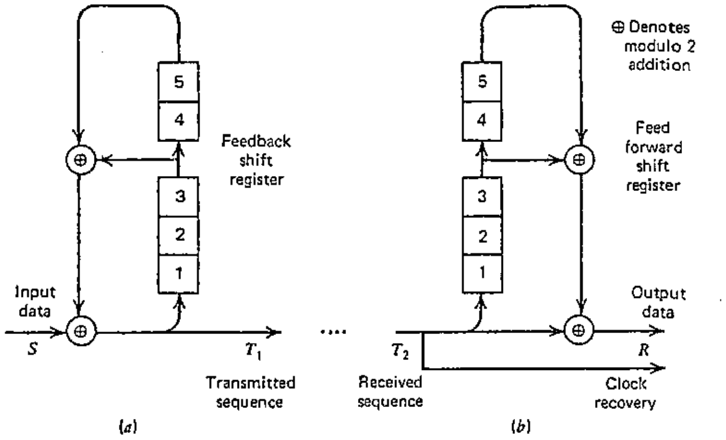


Figure 5.24 (*a*) Scrambler. (*b*) Unscrambler.

In order to analyze the operation of the scrambler and unscrambler, let us introduce an operator "D" to denote the effect of delaying a bit sequence by one bit. Thus DS represents a sequence S delayed by one bit and D^kS represents the sequence S delayed by k bits. Using the delay operator, we can establish the following relationships:

Starting with the unscrambler, we have $R = T_2 \oplus D^3T_2 \oplus D^5T_2 = (1 \oplus F)T_2$, where F stands for the operator $D^3 \oplus D^5$. In the scrambler, T_1 is operated on by $F = D^3 \oplus D^5$ and added to S. That is, $T_1 = S \oplus FT_1$ or

$$T_1 = \frac{S}{1 \oplus F} \tag{5.70}$$

where $F = D^3 \oplus D^5$ and division stands for inverse operator. In the absence of errors, we have $T_2 = T_1$ and hence the unscrambled output

$$\begin{aligned} R &= (1 \oplus F)T_2 \\ &= (1 \oplus F)T_1 = \frac{1 \oplus F}{1 \oplus F}S \\ &= S \end{aligned} \tag{5.71}$$

Thus the input sequence is exactly duplicated at the output of the unscrambler.

To illustrate the effect of the scrambler on periodic sequences and on long strings of ones or zeros, let us consider an input sequence shown in Table 5.3 and assume that the initial content of the register is zeros. This table

Table 5.3. Input and output bit streams of the scrambler shown in Figure 5.24

Input S	1 0 1 0 1 0 1 0 0 0 0 0 1 1
D^3T_1	0 0 0 1 0 1 1 1 0 0 0 1 1 0
D^5T_1	0 0 0 0 0 1 0 1 1 1 0 0 0 1
Output T_1	1 0 1 1 1 0 0 0 1 1 0 1 0 0

illustrates that the scrambler can effectively remove periodic sequences and long strings of zeros by scrambling the data input. The scrambler in general produces a pseudo-random sequence given zeros as a data input, assuming it starts from a nonzero state. With an appropriate choice of taps, an n-bit shift register scrambler can be made to produce a sequence of $2^n - 1$ bits before it repeats itself. The design of the feedback and feed forward registers used in scramblers and unscramblers is rather involved. The interested reader can find a good treatment on the analysis and synthesis of these devices in texts dealing with algebraic coding schemes (see, for example, Peterson's book on coding).

Introduction of scramblers affects the error performance of the communication system in that a single channel error may cause multiple errors at the output of the unscrambler. This is due to the propagation of the error bit in the shift register at the unscrambler. Fortunately, the error propagation effect lasts over only a finite, often small, number of bits. In the scrambler/unscrambler shown in Figure 5.24, each isolated error bit causes three errors in the final output. It must also be pointed out that some random bit patterns might be scrambled to all zeros or all ones.

5.8 SUMMARY

In this chapter we developed procedures for designing and analyzing baseband PAM data transmission systems. The main objectives of the design were to eliminate intersymbol interference and minimize the effects of noise. Several methods of data transmission using PAM techniques were considered. The performance of baseband PAM systems was compared in terms of power-bandwidth requirements for a given data rate and error rate. Methods of shaping the transmitted signal and its spectrum were discussed. The problems of equalization and clock recovery were considered and methods of automatic equalization and synchronization were presented.

REFERENCES

A detailed and thorough treatment of several aspects of baseband system design may be found in the book by Lucky et al. (1968). This book is written for advanced graduate students and the average reader may find it hard to read. Practical aspects of baseband data transmission are dealt with, rather nicely, in the book by Davies and Barber (1975). Introductory level treatment of baseband PAM systems may be found in many undergraduate texts [Bennet and Davey (1965), Carlson (1975), and Ziemer and Tranter (1976)]. Carlson's book contains an easily readable treatment of PAM systems with several examples.

1. R. W. Lucky et al. *Principles of Data Communication.* McGraw-Hill, New York (1968), Chapter 4.
2. D. W. Davies and D. L. A. Barber. *Communication Networks for Computers.* Wiley, New York (1975), Chapter 5.
3. W. R. Bennet and J. R. Davey. *Data Transmission.* McGraw-Hill, New York (1965).
4. A. B. Carlson. *Communication Systems.* McGraw-Hill, New York (1975).
5. R. E. Ziemer and W. H. Tranter. *Principles of Communications.* Houghton Mifflin, Boston (1976).
6. W. W. Peterson. *Error Correcting Codes.* MIT Press, Cambridge, MA (1961).

PROBLEMS

Section 5.1

5.1. A baseband binary PAM system uses a signaling waveform

$$p_g(t) = \frac{\sin \pi r_b t}{\pi r_b t}$$

to transmit binary data at a bit rate of r_b bits/sec. The amplitude levels at the pulse generator output are +1 volt or −1 volt, +1 if the input bit is 1 and −1 if the input bit is 0. Sketch the waveform of the pulse generator output when the input bit string is 0 0 1 0 1 1 0.

5.2. Suppose that the received pulse in a baseband binary PAM system is given by

$$p_r(t) = \frac{\sin \pi r_b t}{\pi r_b t}$$

with amplitude levels ±1 mvolt. The received waveform is sampled at $t_k = kT_b + (T_b/10)$ $(k = 0, \pm 1, \ldots, \pm M)$, that is, there is a timing error of one tenth of a bit duration. Assuming that the input to the system is a sequence of $2M+1$ bits of alternating 0's and 1's, find the value of the ISI term at t_0.

✓ 5.3. Suppose that the received pulse in a baseband binary system has the shape shown in Figure 5.25. Consider the input to the A/D converter

$$Y(t) = \sum_k A_k p_r(t - t_d - kT_b), \quad A_k = \pm 1$$

Assuming $t_d = T_b/2$ and $\tau = 2T_b$, find the value of the ISI term when the input bit sequence is a long string of alternating zeros and ones.

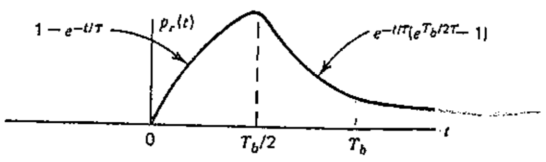


Figure 5.25 $p_r(t)$ for Problem 5.3.

Sections 5.2 and 5.3

✓ 5.4. We want to select a $P_r(f)$ for transmitting binary data at a rate of r_b bits/sec. Which one of the three shown in Figure 5.26 would you choose? Give the reasons for your choice.

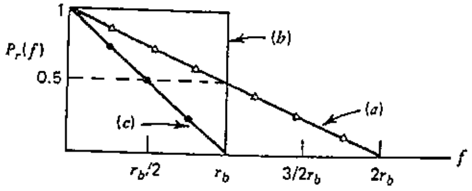


Figure 5.26 $P_r(f)$ for Problem 5.4.

5.5. Derive the result stated in the footnote on page 198.

✓ 5.6. In a binary PAM system, the sampled value of the received waveform Y has the following probability density functions depending on the input bit:

$$f_{Y|1\text{ sent}}(y) = \frac{1}{\sqrt{2\pi}} \exp(-(y-1)^2/2), \quad -\infty < y < \infty$$

$$f_{y|0\text{ sent}}(y) = \frac{1}{\sqrt{2\pi}} \exp(-(y+1)^2/2), \quad -\infty < y < \infty$$

$$P(1\text{ sent}) = p, \quad P(0\text{ sent}) = 1 - p$$

The receiver compares Y against a threshold value T and outputs a 1 if $Y > T$ and a 0 if $Y \leq T$.

(a) Derive an expression for the threshold T that minimizes the probability of incorrectly decoding a bit. Find T for $p = 0.2$, 0.4, 0.5, 0.6, and 0.8.

(b) Calculate the probability of error for each of the above values of p.

✓ 5.7. Calculate P_e for the system of Problem 5.6 with $T = 0$ for $p = 0.2$, 0.4, 0.5, 0.6, and 0.8 and compare with the results obtained in that problem.

✓ 5.8. The sampled values of the received waveform in a binary PAM system suffer from ISI such that

$$Y(t_m) = \begin{cases} A_m + n(t_m) + Q & \text{when the input bit} = 1 \\ A_m + n(t_m) - Q & \text{when the input bit} = 0 \end{cases}$$

where Q is the ISI term. The ISI term has one of three values with the following probabilities:

$$P(Q = +q) = \tfrac{1}{4}$$
$$P(Q = 0) = \tfrac{1}{2}$$
$$P(Q = -q) = \tfrac{1}{4}$$

(a) Assume that $n(t_m)$ is a Gaussian random variable with a variance of σ^2 and that $A_m = +A$ or $-A$ depending on whether the transmitted bit is 1 or 0. Derive an expression for the probability of error in terms of A, σ, and q.

(b) Find P_e for $A/\sigma = 3.0$, and $q/A = 0.1$ and 0.25. How much does the ISI affect the probability of error in each case?

✓ 5.9. Design a binary baseband PAM system to transmit data at a rate of 9600 bits/sec with a bit error probability $P_e < 10^{-5}$. The channel available is an ideal lowpass channel with a bandwidth of 9600 Hz. The noise can be assumed to be white, Gaussian with a two-sided psd $\eta/2 = 10^{-13}$ watt/Hz. Sketch the shape of $|H_T(f)|$, $|H_R(f)|$, $|P_g(f)|$, and find the transmitter power requirements.

✓ 5.10. Repeat Problem 5.9 with

$$H_c(f) = \frac{1}{1 + j(f/f_c)}, \quad f_c = 4800 \text{ Hz}$$

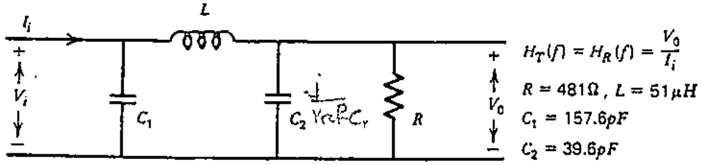

L
I_i
V_i
C_1
C_2
R
V_0
$H_T(f) = H_R(f) = \frac{V_0}{I_i}$
$R = 481\Omega$, $L = 51\mu H$
$C_1 = 157.6pF$
$C_2 = 39.6pF$

Figure 5.27 Filter network for Problem 5.11.

√ 5.11. The filter shown in Figure 5.27 is used for both the transmitting and receiving filter in a binary PAM system. The channel is ideal lowpass with additive Gaussian noise.

(a) Assume that the bit rate at the input is $(6.28)(10^6)$ bits/sec and $p_g(t)$ is a rectangular pulse with a width $= T_b$. Find $p_r(t)$ and sketch it.

(b) Is there any ISI in the received waveform? Can you use this filter for binary data transmission at all?

√ 5.12. The following $P_r(f)$ are used in binary data transmission with controlled ISI:

(a) $4T_b \cos^2 \pi f T_b \begin{cases} & \text{for } |f| \leq \dfrac{1}{2T_b} \\ 0 & \text{elsewhere} \end{cases}$

(b) $2T_b \sin 2\pi f T_b \begin{cases} & \text{for } |f| \leq \dfrac{1}{2T_b} \\ 0 & \text{elsewhere} \end{cases}$

(c) $4T_b \sin^2(2\pi f T_b) \begin{cases} & \text{for } |f| \leq \dfrac{1}{2T_b} \\ 0 & \text{elsewhere} \end{cases}$

For each of the above cases, find $p_r(t)$, and the number of received levels.

5.13. A source emits one of three equiprobable symbols in an independent sequence at a symbol rate of 1000/sec. Design a three-level PAM system to transmit the output of this source over an ideal lowpass channel with additive Gaussian noise having a psd of $\eta/2 = 10^{-14}$ watt/Hz. The symbol error probability has to be maintained at or below 10^{-6}. Specify the power, bandwidth requirements, and $H_T(f)$, $H_R(f)$, and $p_g(t)$.

5.14. The received waveform in a three-level system has the values -1 volt, 0 and $+1$ volt in the absence of noise. The probabilities of these levels are $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$, respectively. The additive noise in the system is Gaussian with a standard deviation of $\frac{1}{4}$ volt.

(a) Find the optimum threshold settings for decoding the levels.

(b) Find the probability of error P_e for the optimum decoding scheme.

√ 5.15. Design a PAM system to transmit the output of a source emitting an equiprobable, independent bit stream at a rate of 10,000 bits/sec over an ideal lowpass channel of width 5000 Hz and additive Gaussian noise with a psd $= 10^{-12}$ watt/Hz. P_e has to be maintained at or below 10^{-4}.

5.16. Calculate the capacity of the discrete channel discussed in Problem 5.15 if the bit transition rate is limited to 10,000/sec.

Section 5.5

5.17. Verify the spectral component $G(f)$ for the twinned binary coding scheme given in Equation (5.60).

5.18. A baseband binary communication system uses a received pulse with a spectrum $P_r(f)$ given by

$$P_r(f) = \begin{cases} T_b \cos^2(\pi f / 2r_b), & |f| < r_b \\ 0, & \text{elsewhere} \end{cases}$$

The channel and noise characteristics dictate a $H_T(f) = \sqrt{[P_r(f)]}$. The system uses a bipolar coding scheme for the pulse amplitudes with

$$p_g(t) = \begin{cases} 1 & \text{for } |t| < T_b/20 \\ 0 & \text{elsewhere} \end{cases}$$

Assuming an independent stream of equiprobable bits at the input to the pulse generator, compute the power spectral density of the transmitting filter output.

Section 5.6

√ 5.19. The unequalized pulse in a PAM system has the following values at sampling times:

$$p_r(kT_b) = p_r(k) = \begin{cases} 0.2, & k = 1 \\ 0.8, & k = 0 \\ 0.2, & k = -1 \end{cases}$$

$$p_r(k) = 0 \quad \text{for } |k| > 1$$

(a) Design a three-tap zero forcing equalizer so that the equalizer output is 1 at $k = 0$ and 0 at $k = \pm 1$.

(b) Calculate $p_{eq}(k)$ for $k = \pm 2, \pm 3$.

5.20. Would a five-tap equalizer for Problem 5.19 yield $p_{eq}(k) = 1$ for $k = 0$ and $p_{eq}(k) = 0$ for $k \neq 0$?

✓5.21. A baseband binary PAM system was designed with the assumption that the channel behaved like an ideal lowpass filter with a bandwidth $B = r_b$ Hz. The channel noise was assumed to be white and the pulse spectrum was chosen to be

$$P_r(f) = \begin{cases} T_b \cos^2(\pi f/2r_b), & \text{for } |f| < r_b \\ 0, & \text{elsewhere} \end{cases}$$

(a) Calculate the design value of $(A^2/N_0)_{\max}$. (Kα)

(b) Suppose that the channel response turned out to be

$$H_c(f) = 1/(1 + jfT_b)$$

and an equalizing filter with $H_{eq}(f) = 1/H_c(f)$ was used at the receiver. Calculate the (A^2/N_0) assuming that the transmitting and receiving filters are the same as before. (Note: The signal power in both (a) and (b) will be the same but the equalizer will change the value of the noise power.)

(c) By what factor should the transmitter power be increased to maintain the same error probability?

5.22. A four-level PAM signaling scheme is used to transmit data over an ideal lowpass channel having a bandwidth B. The additive channel noise is zero mean, Gaussian with a power spectral density of $\eta/2$, and the signal-to-noise ratio at the output of the channel is S/N.

(a) Plot C/B versus S/N (in dB) for this channel (C is the channel capacity).

(b) Develop the discrete channel model for the four-level PAM scheme. Find the rate of information transmission D_t over the discrete channel and sketch D_t/B. Compare your results with the plot shown in Figure 4.14 (Chapter 4). (Assume that the input symbols are equiprobable and occur in independent sequences.)

5.23. Repeat Problem 5.22(b) for $M = 8$, 16, and 32.

6

ANALOG SIGNAL TRANSMISSION

6.1 INTRODUCTION

In Chapter 5 we were primarily concerned with the transmission of messages that consisted of sequences of symbols. Each symbol was chosen from a source alphabet consisting of a finite number of symbols. Corresponding to each symbol a particular electrical waveform was transmitted over the channel. Thus messages were represented by sequences of waveforms, each of which was selected from a finite set of known waveforms. The receiver had the simple task of detecting which one of the finite number of known waveforms was transmitted during each symbol interval.

In contrast, we will now be concerned with the transmission of messages that are continuous (or analog) signals. Each message waveform is chosen from an uncountably infinite number of possible waveforms. For example, in radio or television broadcasting we have an uncountably infinite number of possible messages and the corresponding waveforms are not all known. Such a collection of messages and waveforms can be conveniently modeled by continuous random processes wherein each member function of the random process corresponds to a message waveform. We will use the random signal model in Chapter 7 when we discuss the effects of random noise in analog communication systems. For purposes of analysis, let us define analog signal transmission as the transmission of an arbitrary finite energy lowpass

8

DIGITAL CARRIER MODULATION SCHEMES

8.1 INTRODUCTION

In Chapter 5 we described several methods of transmitting digital information over baseband channels using discrete baseband PAM techniques. Most real communication channels have very poor response in the neighborhood of zero frequency and hence are regarded as bandpass channels. In order to transmit digital information over bandpass channels, we have to transfer the information to a carrier wave of appropriate frequency. Digital information can be impressed upon a carrier wave in many different ways. In this chapter, we will study some of the most commonly used digital modulation techniques wherein the digital information modifies the amplitude, the phase, or the frequency of the carrier in discrete steps.

Figure 8.1 shows four different modulation waveforms for transmitting binary information over bandpass channels. The waveform shown in Figure 8.1*a* corresponds to discrete amplitude modulation or an amplitude-shift keying (ASK) scheme where the amplitude of the carrier is switched between two values (on and off). The resultant waveform consists of "on" (mark) pulses representing binary 1 and "off" (space) pulses representing binary 0. The waveform shown in Figure 8.1*b* is generated by switching the frequency of the carrier between two values corresponding to the binary information to be transmitted. This method, where the frequency of the carrier is changed, is

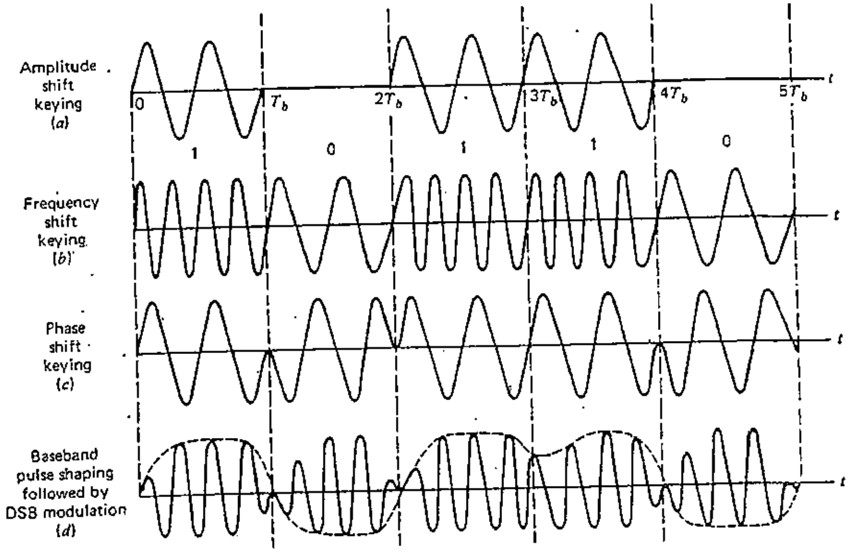

Amplitude shift keying (*a*)
0
T_b
$2T_b$
$3T_b$
$4T_b$
$5T_b$
t
1
0
1
1
0
Frequency shift keying (*b*)
t
Phase shift keying (*c*)
t
Baseband pulse shaping followed by DSB modulation (*d*)
t

Figure 8.1 Modulated carrier waveforms used in binary data transmission schemes. (*a*) Amplitude-shift keying. (*b*) Frequency-shift keying. (*c*) Phase-shift keying. (*d*) Baseband pulse shaping followed by DSB modulation.

called frequency-shift keying (FSK). In the third method of digital modulation shown in Figure 8.1*c*, the carrier phase is shifted between two values and hence this method is called phase-shift keying (PSK). It should be noted here that in PSK and FSK methods, the amplitude of the carrier remains constant. Further, in all cases the modulated waveforms are continuous at all times. Finally, Figure 8.1*d* shows a modulation waveform generated by amplitude modulating the carrier with a baseband signal generated by the discrete PAM scheme described in the previous chapter.

The modulation scheme using baseband pulse shaping followed by analog modulation (DSB or VSB) requires the minimum transmission bandwidth. However, the equipment required to generate, transmit, and demodulate the waveform shown in Figure 8.1*d* is quite complex. In contrast, the digital modulation schemes are extremely simple to implement. The price paid for this simplicity is excessive bandwidth and possible increase in transmitter power requirements. When bandwidth is not the major consideration, then digital modulation schemes provide very good performance with minimum equipment complexity and with a good degree of immunity to certain channel impairments. In the following sections we will study digital modulation schemes. Primary emphasis will be given to the study of system performance in the presence of additive noise as measured by the probability of error.

We will begin our study of digital modulation schemes with the derivation of an optimum receiver for demodulating ASK, PSK, and FSK signals with minimum probability of error. We will show that such a receiver consists of a matched filter (or a correlation receiver) if the additive noise is white. We will derive expressions for the probability of error for various modulation schemes in terms of the average signal power at the receiver input, power spectral density of the noise at the receiver input, and the signaling rate.

In Section 8.3 we will study the amplitude shift-keying (ASK) method. We will look at optimum and suboptimum methods of demodulating binary ASK signals. In Sections 8.4 and 8.5, we will deal with optimum and suboptimum binary PSK and FSK schemes. Finally, in Section 8.6 we will compare the performance of these binary schemes in terms of power-bandwidth requirements and probability of error. In Section 8.7 we will discuss M-ary PSK, DPSK, and FSK schemes. The problem of synchronizing the receiver to the transmitter will be considered in Section 8.8.

8.2 OPTIMUM RECEIVER FOR BINARY DIGITAL MODULATION SCHEMES

The function of a receiver in a binary communication system is to distinguish between two transmitted signals $s_1(t)$ and $s_2(t)$ in the presence of noise. The performance of the receiver is usually measured in terms of the probability of error and the receiver is said to be optimum if it yields the minimum probability of error. In this section, we will derive the structure of an optimum receiver that can be used for demodulating binary ASK, PSK, and FSK signals.

We will show that the optimum receiver takes the form of a matched filter when the noise at the receiver input is white. We will also show that the matched filter can be implemented as an integrate and dump correlation receiver. The integrate and dump correlation receiver is a *coherent* or synchronous receiver that requires a local carrier reference having the same phase and frequency as the transmitted carrier. Elaborate circuitry is required at the receiver to generate the coherent local carrier reference.

The binary ASK, PSK, and FSK signals can also be demodulated using suboptimal *noncoherent* demodulation schemes. Such schemes are simpler to implement and are widely used in low speed data transmission applications. We will deal with suboptimal (noncoherent) methods of demodulating binary ASK, PSK, and FSK signals in Sections 8.3, 8.4, and 8.5.

8.2.1 Description of Binary ASK, PSK, and FSK Schemes

The block diagram of a bandpass binary data transmission scheme using digital modulation is shown in Figure 8.2. The input to the system is a binary bit sequence $\{b_k\}$ with a bit rate r_b and bit duration T_b. The output of the modulator during the kth bit interval depends on the kth input bit b_k. The modulator output $Z(t)$ during the kth bit interval is a shifted version of one of two basic waveforms $s_1(t)$ or $s_2(t)$, and $Z(t)$ is a random process defined by

$$Z(t) = \begin{cases} s_1[t-(k-1)T_b] & \text{if } b_k = 0 \\ s_2[t-(k-1)T_b] & \text{if } b_k = 1 \end{cases} \tag{8.1}$$

for $(k-1)T_b \le t \le kT_b$. The waveforms $s_1(t)$ and $s_2(t)$ have a duration of T_b and have finite energy, that is, $s_1(t)$ and $s_2(t) = 0$ if $t \notin [0, T_b]$ and

$$\begin{aligned} E_1 &= \int_0^{T_b} [s_1(t)]^2\,dt < \infty \\ E_2 &= \int_0^{T_b} [s_2(t)]^2\,dt < \infty \end{aligned} \tag{8.2}$$

The shape of the waveforms depends on the type of modulation used, as shown in Table 8.1. The output of the modulator passes through a bandpass channel $H_c(f)$, which, for purposes of analysis, is assumed to be an ideal channel with adequate bandwidth so that the signal passes through without suffering any distortion other than propagation delay. The channel noise $n(t)$ is assumed to be a zero mean, stationary, Gaussian random process with a known power spectral density $G_n(f)$. The received signal plus noise then is

$$V(t) = \begin{cases} s_1[t-(k-1)T_b - t_d] + n(t) \\ \quad\text{or} \\ s_2[t-(k-1)T_b - t_d] + n(t) \end{cases} \qquad (k-1)T_b + t_d \le t \le kT_b + t_d$$

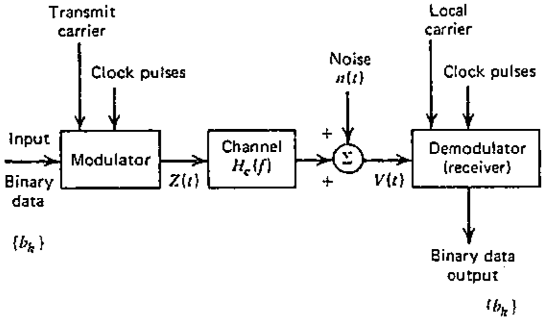


Figure 8.2 Bandpass binary data transmission system.

Table 8.1. Choice of signaling waveforms for various types of digital modulation schemes. $s_1(t)$, $s_2(t) = 0$ for $t \notin [0, T_b]$; $f_c = \omega_c/2\pi$. The frequency of the carrier f_c is assumed to be a multiple of r_b.

$s_1(t)$; $0 \leq t \leq T_b$	$s_2(t)$; $0 \leq t \leq T_b$	Type of Modulation
0	$A \cos \omega_c t$ (or $A \sin \omega_c t$)	Amplitude-shift keying (ASK)
$-A \cos \omega_c t$ (or $-A \sin \omega_c t$)	$A \cos \omega_c t$ ($A \sin \omega_c t$)	Phase-shift keying (PSK)
$A \cos\{(\omega_c - \omega_d)t\}$ (or $A \sin\{(\omega_c - \omega_d)t\}$)	$A \cos\{(\omega_c + \omega_d)t\}$ ($A \sin\{(\omega_c + \omega_d)t\}$)	Frequency-shift keying (FSK)

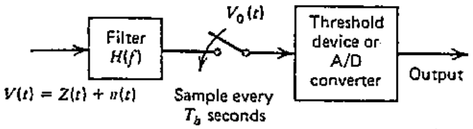


Figure 8.3 Receiver structure.

where t_d is the propagation delay, which can be assumed to be zero without loss of generality.

The receiver shown in Figure 8.3 has to determine which of the two *known waveforms* $s_1(t)$ or $s_2(t)$ was present at its input during each signaling interval. The actual receiver consists of a filter, a sampler, and a threshold device. The signal plus noise $V(t)$ is filtered and sampled at the end of each bit interval. The sampled value is compared against a predetermined threshold value T_0 and the transmitted bit is decoded (with occasional errors) as 1 or 0 depending on whether $V_0(kT_b)$ is greater or less than the threshold T_0.

The receiver makes errors in the decoding process due to the noise present at its input. The error probability will depend on the signal power at the receiver input, noise power spectral density at the input, signaling rate, and receiver parameters such as the filter transfer function $H(f)$ and threshold setting.

8.2.2 Probability of Error

The measure of performance used for comparing digital modulation schemes is the probability of error. The receiver parameters such as $H(f)$ and threshold setting are chosen to minimize the probability of error. In this section, we will derive expressions for the probability of error in terms of the signal parameters, noise power spectral density, and the receiver parameters. These expressions will be used in the following sections for the analysis and design of digital modulation schemes.

We will make the following assumptions while deriving the expressions for the probability of error:

1. We will assume that $\{b_k\}$ is an equiprobable, independent sequence of bits. Hence, the occurrence of $s_1(t)$ or $s_2(t)$ during a bit interval does not influence the occurrence of $s_1(t)$ or $s_2(t)$ during any other non-overlapping bit interval; further, $s_1(t)$ and $s_2(t)$ are equiprobable.
2. The channel noise will be assumed to be a zero mean Gaussian random process with a power spectral density $G_n(f)$.
3. We will assume that the intersymbol interference generated by the filter is small.*

The output of the filter at $t = kT_b$ can be written as

$$V_0(kT_b) = s_0(kT_b) + n_0(kT_b) \tag{8.3}$$

where $s_0(t)$ and $n_0(t)$ denote the response of the filter due to signal and noise components at its input. The signal component in the output at $t = kT_b$ is given by

$$s_0(kT_b) = \int_{-\infty}^{kT_b} Z(\zeta)h(kT_b - \zeta)\,d\zeta \tag{8.4}$$

$$= \int_{(k-1)T_b}^{kT_b} Z(\zeta)h(kT_b - \zeta)\,d\zeta + \text{ISI terms} \tag{8.5}$$

where $h(\zeta)$ is the impulse response of the filter. Since we have assumed the ISI terms to be zero, we can rewrite Equation (8.5) as

$$s_0(kT_b) = \int_{(k-1)T_b}^{kT_b} Z(\zeta)h(kT_b - \zeta)\,d\zeta$$

Substituting $Z(t)$ from Equation (8.1) and making a change of variable, we can write the signal component as

$$s_0(kT_b) = \begin{cases} \displaystyle\int_0^{T_b} s_1(\zeta)h(T_b - \zeta)\,d\zeta = s_{01}(kT_b) \text{ when } b_k = 0 \\ \displaystyle\int_0^{T_b} s_2(\zeta)h(T_b - \zeta)\,d\zeta = s_{02}(kT_b) \text{ when } b_k = 1 \end{cases} \tag{8.6}$$

*We will see later that the optimum filter for the white noise case generates zero ISI. For colored noise case the optimum filter generates nonzero ISI, which can be minimized by making $s_1(t)$ and $s_2(t)$ to have a duration $\ll T_b$ so that the filter response settles down to a negligible value before the end of each bit interval.

The noise component $n_0(kT_b)$ is given by

$$n_0(kT_b) = \int_{-\infty}^{kT_b} n(\zeta)h(kT_b - \zeta)\, d\zeta \tag{8.7}$$

The output noise $n_0(t)$ is a stationary zero mean Gaussian random process. The variance of $n_0(t)$ is

$$N_0 = E\{n_0^2(t)\} = \int_{-\infty}^{\infty} G_n(f)|H(f)|^2\, df \tag{8.8}$$

and the probability density function of $n_0(t)$ is

$$f_{n_0}(n) = \frac{1}{\sqrt{2\pi N_0}} \exp\left(\frac{-n^2}{2N_0}\right), \quad -\infty < n < \infty \tag{8.9}$$

The receiver decodes the kth bit by comparing $V_0(kT_b)$ against the threshold T_0. If we assume that $s_1(t)$ and $s_2(t)$ are chosen such that $s_{01}(T_b) < s_{02}(T_b)$, and that the receiver decodes the kth bit as 0 if $V_0(kT_b) < T_0$ and as 1 if $V_0(kT_b) \geqslant T_0$, then the probability that the kth bit is incorrectly decoded is given by P_e, where

$$\begin{aligned} P_e &= P\{b_k = 0 \text{ and } V_0(kT_b) \geqslant T_0;\text{ or } b_k = 1 \text{ and } V_0(kT_b) < T_0\} \\ &= \tfrac{1}{2}P\{V_0(kT_b) \geqslant T_0 | b_k = 0\} \\ &\quad + \tfrac{1}{2}P\{V_0(kT_b) < T_0 | b_k = 1\} \end{aligned} \tag{8.10}$$

If the kth transmitted bit is 0, then $V_0 = s_{01} + n_0$ where s_{01} is a constant and n_0 is a zero mean Gaussian random variable with the variance given in Equation (8.8). Hence, the conditional pdf of V_0 given $b_k = 0$ is given by

$$f_{V_0|b_k=0}(v_0) = \frac{1}{\sqrt{2\pi N_0}} \exp\left(\frac{-(v_0 - s_{01})^2}{2N_0}\right), \quad -\infty < v_0 < \infty \tag{8.11a}$$

Similarly, when b_k is 1, the conditional pdf of V_0 has the form

$$f_{V_0|b_k=1}(v_0) = \frac{1}{\sqrt{2\pi N_0}} \exp\left\{\frac{-(v_0 - s_{02})^2}{2N_0}\right), \quad -\infty < v_0 < \infty \tag{8.11b}$$

Combining Equations (8.10) and (8.11), we obtain an expression for the probability of error P_e as

$$\begin{aligned} P_e &= \tfrac{1}{2}\int_{T_0}^{\infty} \frac{1}{\sqrt{2\pi N_0}} \exp\left(\frac{-(v_0 - s_{01})^2}{2N_0}\right) dv_0 \\ &\quad + \tfrac{1}{2}\int_{-\infty}^{T_0} \frac{1}{\sqrt{2\pi N_0}} \exp\left(\frac{-(v_0 - s_{02})^2}{2N_0}\right) dv_0 \end{aligned} \tag{8.12}$$

Because of equal probabilities of occurrences of 0's and 1's in the input bit stream and the symmetrical shapes of $f_{V_0|b_k=0}$ and $f_{V_0|b_k=1}$ shown in Figure 8.4, it can be shown that the optimum choice for the threshold is the value of v_0 at which the conditional pdf's intersect (see Problem 8.5). This optimum value of the threshold T_0^* is

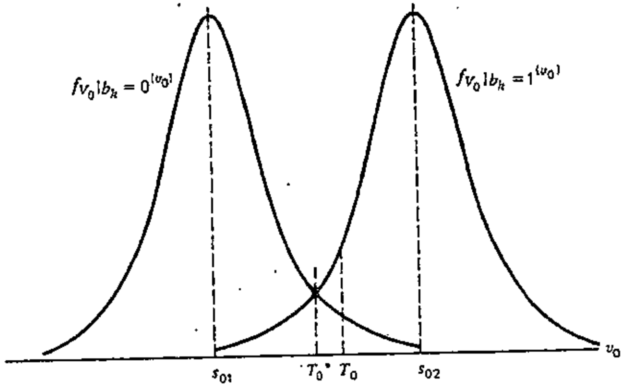


Figure 8.4 Conditional pdf of V_0 given b_k.

$$T_0^* = \frac{s_{01} + s_{02}}{2}$$

Substituting the value of T_0^* for T_0 in (8.12), we can rewrite the expression for the probability of error as

$$\begin{aligned} P_e &= \int_{(s_{02}+s_{01})/2}^{\infty} \frac{1}{\sqrt{2\pi N_0}} \exp\left(-\frac{(v_0 - s_{01})^2}{2N_0}\right) dv_0 \\ &= \int_{(s_{02}-s_{01})/2\sqrt{N_0}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz \end{aligned} \tag{8.13}$$

The above expression for the probability of error is a monotonically decreasing function in its argument, that is, P_e becomes smaller as $(s_{02} - s_{01})/\sqrt{N_0}$ becomes larger. Equations (8.6), (8.7) and (8.8) indicate that s_{01}, s_{02}, and $\sqrt{N_0}$ depend on the choice of the filter impulse response (or the transfer function). The optimum filter then is the filter that maximizes the ratio

$$\gamma = \frac{s_{02}(T_b) - s_{01}(T_b)}{\sqrt{N_0}} \tag{8.14}$$

or the square of the ratio γ^2. Observe that maximizing γ^2 eliminates the requirement $s_{01} < s_{02}$.

8.2.3 Transfer Function of the Optimum Filter

The essential function of the receiver shown in Figure 8.3 is that it has to determine which of the two *known waveforms* $s_1(t)$ or $s_2(t)$ was present at its input during each signaling interval. The optimum receiver distinguishes between $s_1(t)$ and $s_2(t)$ from the noisy versions of $s_1(t)$ and $s_2(t)$ with minimum probability of error. We have seen in the preceding section that the probability of error is minimized by an appropriate choice of $h(t)$ which maximizes γ^2, where

$$\gamma^2 = \frac{[s_{02}(T_b) - s_{01}(T_b)]^2}{N_0} \tag{8.15}$$

$$s_{02}(T_b) - s_{01}(T_b) = \int_0^{T_b} [s_2(\zeta) - s_1(\zeta)]h(T_b - \zeta)\,d\zeta$$

and

$$N_0 = \int_{-\infty}^{\infty} G_n(f)|H(f)|^2\,df$$

If we let $p(t) = s_2(t) - s_1(t)$, then the numerator of the quantity to be maximized is

$$\begin{aligned} s_{02}(T_b) - s_{01}(T_b) = p_0(T_b) &= \int_0^{T_b} p(\zeta)h(T_b - \zeta)\,d\zeta \\ &= \int_{-\infty}^{\infty} p(\zeta)h(T_b - \zeta)\,d\zeta \end{aligned} \tag{8.16}$$

since $p(t) = 0$ for $t < 0$ and $h(\lambda) = 0$ for $\lambda < 0$.
If we let $P(f)$ be the Fourier transform of $p(t)$, then we can obtain the Fourier transform $P_0(f)$ of $p_0(t)$ from Equation (8.16) as

$$P_0(f) = P(f)H(f)$$

or

$$p_0(T_b) = \int_{-\infty}^{\infty} P(f)H(f)\exp(j2\pi fT_b)\,df$$

Hence γ^2 can be written as

$$\gamma^2 = \frac{\left|\int_{-\infty}^{\infty} H(f)P(f)\exp(j2\pi fT_b)\,df\right|^2}{\int_{-\infty}^{\infty} |H(f)|^2 G_n(f)\,df} \tag{8.17}$$

We can maximize γ^2 by applying Schwarz's inequality, which has the form

$$\frac{\left|\int_{-\infty}^{\infty} X_1(f)X_2(f)\,df\right|^2}{\int_{-\infty}^{\infty} |X_1(f)|^2\,df} \le \int_{-\infty}^{\infty} |X_2(f)|^2\,df \tag{8.18}$$

where $X_1(f)$ and $X_2(f)$ are arbitrary complex functions of a common variable f. The equal sign in (8.18) applies when $X_1(f) = KX_2^*(f)$, where K is an arbitrary constant and $X_2^*(f)$ is the complex conjugate of $X_2(f)$. Applying Schwarz's inequality to Equation (8.17) with

$$X_1(f) = H(f)\sqrt{G_n(f)}$$

and

$$X_2(f) = \frac{P(f)\exp(j2\pi fT_b)}{\sqrt{G_n(f)}}$$

we see that $H(f)$, which maximizes γ^2, is given by

$$H(f) = K\frac{P^*(f)\exp(-j2\pi fT_b)}{G_n(f)} \tag{8.19}$$

where K is an arbitrary constant. Substituting Equation (8.19) in (8.17), we obtain the maximum value of γ^2 as

$$\gamma_{\max}^2 = \int_{-\infty}^{\infty} \frac{|P(f)|^2}{G_n(f)}\,df \tag{8.20}$$

and the minimum probability of error is given by

$$\begin{aligned} P_e &= \int_{\gamma_{\max}/2}^{\infty} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right)dz \\ &= Q\left(\frac{\gamma_{\max}}{2}\right) \end{aligned} \tag{8.21}$$

Special Case I: Matched Filter Receiver. If the channel noise is white, that is, $G_n(f) = \eta/2$, then the transfer function of the optimum receiver is given by

$$H(f) = P^*(f)\exp(-j2\pi fT_b) \tag{8.22}$$

(from Equation (8.19) with the arbitrary constant K set equal to $\eta/2$). The impulse response of the optimum filter is

$$h(t) = \int_{-\infty}^{\infty} [P^*(f)\exp(-2\pi jfT_b)]\exp(2\pi jft)\,df \tag{8.23}$$

Recognizing the fact that the inverse Fourier transform of $P^*(f)$ is $p(-t)$ and that $\exp(-2\pi jfT_b)$ represents a delay of T_b, we obtain $h(t)$ as

$$h(t) = p(T_b - t)$$

Since $p(t) = s_2(t) - s_1(t)$, we have

$$h(t) = s_2(T_b - t) - s_1(T_b - t) \tag{8.24}$$

The impulse response $h(t)$ in Equation (8.24) is matched to the signal $s_1(t)$ and $s_2(t)$ and for this reason the filter is called a *matched filter.* An example is shown in Figure 8.5 to illustrate the significance of the result stated in Equation (8.24).

Figures 8.5*a* and 8.5*b* show $s_2(t)$ and $s_1(t)$ of duration T_b. The waveform $p(t) = s_2(t) - s_1(t)$ is shown in Figure 8.5*c* and $p(-t)$, which is the waveform $p(t)$ reflected around $t = 0$ is shown in 8.5*d.* Finally, the impulse response of the filter $h(t) = p(T_b - t)$, which is $p(-t)$ translated in the positive t direction by T_b, is shown in Figure 8.5*e.* We note here that the filter is causal ($h(t) = 0$ for $t < 0$) and the impulse response has a duration of T_b. The last fact ensures that the signal component of the output at the end of the kth-bit interval is due to signal component at the input during the kth-bit interval only. Thus, there is no intersymbol interference. The probability of error for the matched filter receiver can be obtained from Equations (8.20) and (8.21).

In general, it is very hard to synthesize physically realizable filters that would closely approximate the transfer function specified in Equation (8.22). In the following section we will derive an alternate form for the matched filter that is easier to implement using very simple circuitry.

Special Case II: Correlation Receiver. We will now derive a form of the optimum receiver, which is different from the matched filter implementation. We start with the output of the receiver at $t = T_b$,

$$V_0(T_b) = \int_{-\infty}^{T_b} V(\zeta)h(T_b - \zeta)\, d\zeta$$

where $V(\zeta)$ is the noisy input to the receiver. Substituting $h(\zeta) = s_2(T_b - \zeta) - s_1(T_b - \zeta)$ and noting that $h(\zeta) = 0$ for $\zeta \not\in (0, T_b)$, we can rewrite the preceding expression as

$$\begin{aligned} V_0(T_b) &= \int_0^{T_b} V(\zeta)[s_2(\zeta) - s_1(\zeta)]\, d\zeta \\ &= \int_0^{T_b} V(\zeta)s_2(\zeta)\, d\zeta - \int_0^{T_b} V(\zeta)s_1(\zeta)\, d\zeta \end{aligned} \tag{8.25}$$

Equation (8.25) suggests that the optimum receiver can be implemented as shown in Figure 8.6. This form of the receiver is called a *correlation receiver.*

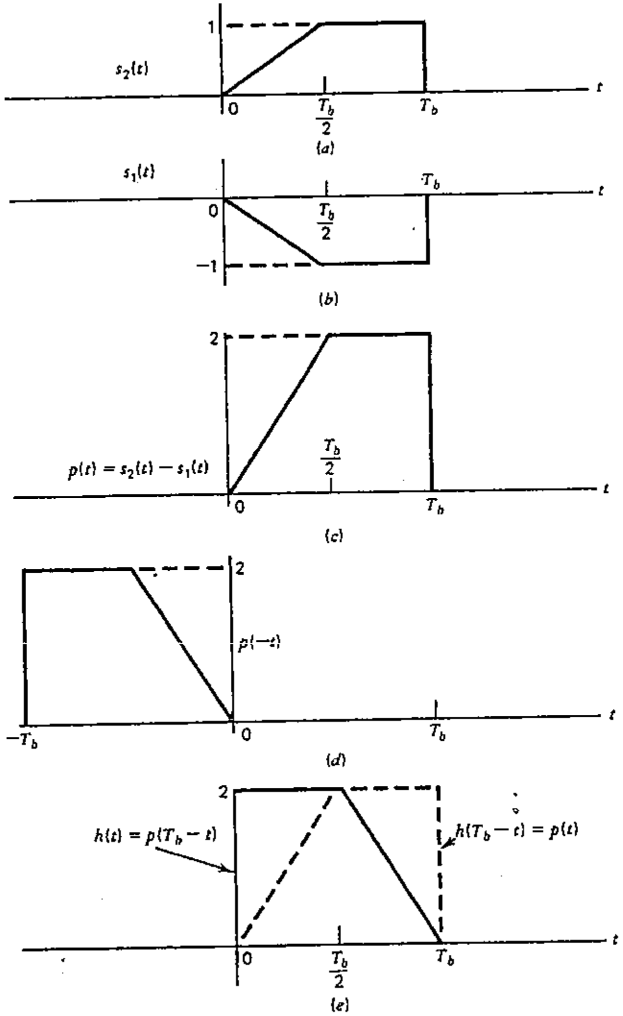


Figure 8.5 Impulse response of a matched filter. (*a*) $s_2(t)$. (*b*) $s_1(t)$. (*c*) $p(t) = s_2(t) - s_1(t)$. (*d*) $p(-t)$. (*e*) $h(t) = p(T_b - t)$.

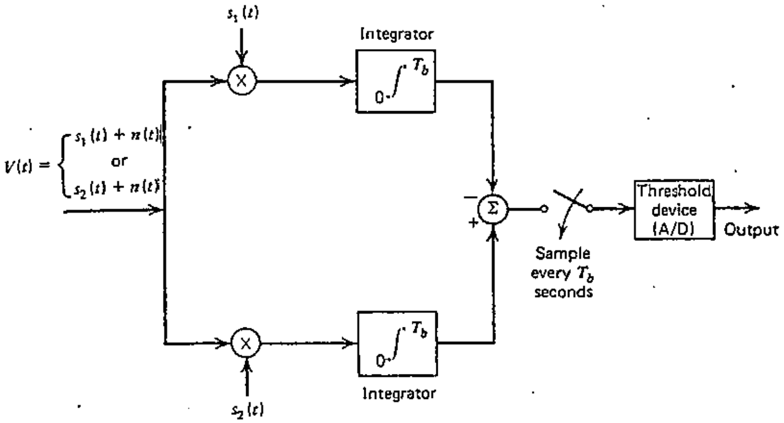

$s_1(t)$
Integrator
$\int_0^{T_b}$
$V(t) = \begin{cases} s_1(t) + n(t) \\ \text{or} \\ s_2(t) + n(t) \end{cases}$
Σ
−
+
Threshold device (A/D)
Output
Sample every T_b seconds
$\int_0^{T_b}$
Integrator
$s_2(t)$

Figure 8.6 Correlation receiver.

It must be pointed out that Equation (8.25) and the receiver shown in Figure 8.6 require that the integration operation be ideal with zero initial conditions. In actual practice, the receiver shown in Figure 8.6 is actually implemented as shown in Figure 8.7. In this implementation, the integrator has to be reset (i.e., the capacitor has to be discharged or dumped) at the end of each signaling interval in order to avoid intersymbol interference. If $RC \gg T_b$, the circuit shown in Figure 8.7 very closely approximates an ideal integrator and operates with the same probability of error as the ideal receiver shown in Figure 8.6.

Needless to say, the sampling and discharging of the capacitor (dumping) must be carefully synchronized. Furthermore, the *local reference signal* $s_2(t) - s_1(t)$ must be in "phase" with the signal component at the receiver input, that is, the correlation receiver performs *coherent* demodulation. The correlation receiver, also known as an *integrate and dump filter*, represents one of the few cases where matched filtering is closely approximated in practice.

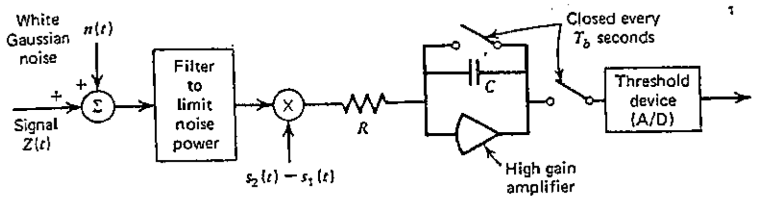


Figure 8.7 Integrate and dump correlation receiver. The bandwidth of the filter preceding the integrator is assumed to be wide enough to pass $Z(t)$ without distortion.

Example 8.1. A bandpass data transmission scheme uses a PSK signaling scheme with

$$s_2(t) = A\cos\omega_c t, \quad 0 \leqslant t \leqslant T_b, \quad \omega_c = 10\pi/T_b$$

$$s_1(t) = -A\cos\omega_c t, \quad 0 \leqslant t \leqslant T_b, \quad T_b = 0.2\ \text{msec}$$

The carrier amplitude at the receiver input is 1 mvolt and the psd of the additive white Gaussian noise at the input is 10^{-11} watt/Hz. Assume that an ideal correlation receiver is used. Calculate the average bit error rate of the receiver.

Solution

$$\text{Data rate} = 5000\ \text{bits/sec}, \quad G_n(f) = \eta/2 = 10^{-11}\ \text{watt/Hz}$$

$$\begin{aligned}\text{Receiver impulse response} &= h(t)\\ &= s_2(T_b - t) - s_1(T_b - t)\\ &= 2A\cos\omega_c(T_b - t)\end{aligned}$$

Threshold setting is 0 and

$$\begin{aligned}\gamma_{\max}^2 &= \int_{-\infty}^{\infty} \frac{|P(f)|^2}{G_n(f)}\,df \text{—(from Equation (8.20))}\\ &= \left(\frac{2}{\eta}\right)\int_{-\infty}^{\infty} |P(f)|^2\,df\\ &= \left(\frac{2}{\eta}\right)\int_0^{T_b} [s_2(t) - s_1(t)]^2\,dt \quad \text{(by Parseval's theorem)}\\ &= \left(\frac{2}{\eta}\right)\int_0^{T_b} 4A^2(\cos\omega_c t)^2\,dt\\ &= \left(\frac{2}{\eta}\right)(2A^2T_b) = \frac{4A^2T_b}{\eta} = 40\end{aligned}$$

$$\begin{aligned}\text{Probability of error} = P_e &= \int_{\frac{1}{2}\gamma_{\max}}^{\infty} \frac{1}{\sqrt{2\pi}}\exp\left(\frac{-z^2}{2}\right)dz\\ &= Q(\sqrt{10})\end{aligned}$$

From the table of Gaussian probabilities, we get $P_e \simeq 0.0008$ and

$$\text{Average error rate} \simeq (r_b)P_e/\text{sec} = 4\ \text{bits/sec}$$

8.3 BINARY ASK SIGNALING SCHEMES

The binary ASK signaling scheme was one of the earliest forms of digital modulation used in wireless (radio) telegraphy at the beginning of this century. While amplitude-shift keying is no longer widely used in digital communications, for reasons that will be discussed later, it is the simplest form of digital modulation and serves as a useful model for introducing certain concepts. The binary ASK waveform can be described as

$$Z(t) = \begin{cases} s_1[t-(k-1)T_b] & \text{if } b_k = 0 \\ s_2[t-(k-1)T_b] & \text{if } b_k = 1 \end{cases} \quad (k-1)T_b \leq t \leq kT_b$$

where $s_2(t) = A\cos\omega_c t$ $(0 \leq t \leq T_b)$ and $s_1(t) = 0$. We will assume that the carrier frequency $\omega_c = 2n\pi/T_b$, where n is an integer.

We can represent $Z(t)$ as

$$Z(t) = D(t)(A\cos\omega_c t) \tag{8.26}$$

where $D(t)$ is a lowpass pulse waveform consisting of (often but not necessarily) rectangular pulses. For purposes of analysis we will assume that $D(t)$ is a rectangular random binary waveform with bit duration T_b. The model for $D(t)$ is (Chapter 3, Section 3.5)

$$d(t) = \sum_{k=-\infty}^{\infty} b_k g[t-(k-1)T_b], \quad b_k = 0 \text{ or } 1$$

$$g(t) = \begin{cases} 1 & 0 \leq t \leq T_b \\ 0 & \text{elsewhere} \end{cases} \tag{8.27}$$

$$D(t) = d(t-T)$$

where T represents a random delay with a uniform pdf in the interval $[0, T_b]$. The form of the modulated waveform $Z(t)$ suggests that the ASK signal can be generated by product modulation, that is, by multiplying the carrier with the rectangular waveform $D(t)$ or using $D(t)$ to turn the carrier oscillator on and off.

The bandwidth requirements for transmitting and processing the ASK signal can be obtained from the power spectral density of $Z(t)$, which can be computed as follows: From Equation (8.26) we see that the power spectral density $G_Z(f)$ of $Z(t)$ is related to the power spectral density $G_D(f)$ of $D(t)$ by*

$$G_Z(f) = \frac{A^2}{4}[G_D(f-f_c) + G_D(f+f_c)] \tag{8.28}$$

*Strictly speaking, we need to include a random phase for the carrier in Equation (8.26) so that $Z(t)$ is a stationary random process.

The waveform $D(t)$ is a random binary waveform with levels 0 and 1. The autocorrelation function and the power spectral density of $D(t)$ are (from Example 3.9 and Equations (3.68) and (3.69))

$$R_{DD}(\zeta) = \begin{cases} \frac{1}{4} + \frac{T_b - |\zeta|}{4T_b} & \text{for } |\zeta| \leq T_b \\ 0 \text{ for } |\zeta| > T_b \end{cases}$$

$$G_D(f) = \tfrac{1}{4}\left(\delta(f) + \frac{\sin^2 \pi f T_b}{\pi^2 f^2 T_b}\right) \tag{8.29}$$

Substituting Equation (8.29) into (8.28), we obtain the psd of $Z(t)$ as

$$G_Z(f) = \frac{A^2}{16}\left(\delta(f-f_c) + \delta(f+f_c) + \frac{\sin^2 \pi T_b(f-f_c)}{\pi^2 T_b(f-f_c)^2} + \frac{\sin^2 \pi T_b(f+f_c)}{\pi^2 T_b(f+f_c)^2}\right) \tag{8.30}$$

A sketch of $G_Z(f)$, shown in Figure 8.8, indicates that $Z(t)$ is an infinite bandwidth signal. However, for practical purposes, the bandwidth of $Z(t)$ is often defined as the bandwidth of an ideal bandpass filter centered at f_c whose output (with $Z(t)$ as its input) contains, say, 95% of the total average power content of $Z(t)$. It can be shown that such a bandwidth would be approximately $3r_b$ Hz for the ASK signal.

The bandwidth of the ASK signal can be reduced by using smoothed versions of the pulse waveform $D(t)$ instead of rectangular pulse waveforms. For example, if we use a pulse waveform $D(t)$ in which the individual pulses $g(t)$ have the shape,

$$g(t) = \begin{cases} (a/2)[1+\cos(2\pi r_b t - \pi)], & 0 \leq t < T_b \\ 0 & \text{elsewhere} \end{cases}$$

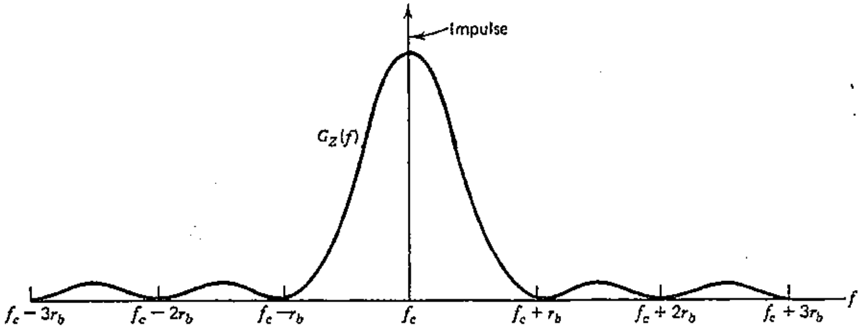


Figure 8.8 Power spectral density of the random binary ASK signal; $f_c \gg r_b$.

$f_c = nT_b^{-1}$

the effective bandwidth of the ASK signal will be of the order of $2r_b$. (The magnitude of the Fourier transform of $g(t)$ drops off as $1/f^3$.) Of course, the gain in bandwidth is somewhat offset by the complexity of the pulse shaping networks needed to generate $g(t)$ given above. Depending on the shape of the pulse waveform, we need a channel with a bandwidth of approximately $2r_b$ to $3r_b$ Hz to transmit an ASK signal.

The transmitted bit sequence $\{b_k\}$ can be recovered from the noisy version of $Z(t)$ at the receiver in one of two ways. The first method of demodulation we will study is the integrate and dump-type coherent demodulation; the second method is the noncoherent envelope demodulation procedure. The principal reason for using the ASK signaling method is its simplicity. Hence, coherent demodulation is seldom used in conjunction with ASK schemes because of the complex circuits needed for maintaining phase coherence between the transmitted signal and the local carrier. Nevertheless, we will investigate the performance of coherent ASK schemes for comparison purposes.

8.3.1 Coherent ASK

The receiver shown in Figure 8.7 can be used for coherent demodulation of an ASK signal. As before, we will assume that the input to the receiver consists of an ASK signal that is corrupted by additive, Gaussian white noise. The receiver integrates the product of the signal plus noise and a copy of the noise free signal over one signaling interval. We will assume that the local signal $s_2(t) - s_1(t) = A \cos \omega_c t$ is carefully synchronized with the frequency and phase of the received carrier. The output of the integrator is compared against a set threshold and at the end of each signaling interval the receiver makes a decision about which of the two signals $s_1(t)$ or $s_2(t)$ was present at its input during the signaling interval. Of course, errors will occur in the demodulation process because of the noise. We can derive an expression for the probability of incorrectly decoding the input waveform using the results derived in Section 8.2.

We start with $s_2(t) = A \cos \omega_c t$, $s_1(t) = 0$, and $s_2(t) - s_1(t) = A \cos \omega_c t$. The signal components of the receiver output at the end of a signaling interval are

$$s_{01}(kT_b) = \int_0^{T_b} s_1(t)[s_2(t) - s_1(t)]\, dt = 0$$

and

$$s_{02}(kT_b) = \int_0^{T_b} s_2(t)[s_2(t) - s_1(t)]\, dt$$
$$= \frac{A^2}{2} T_b$$

In the preceding step, we made use of our assumption that $\omega_c T_b = 2n\pi$, n a positive integer. The optimum threshold setting in the receiver is

$$T_0^* = \frac{s_{01}(kT_b) + s_{02}(kT_b)}{2} = \frac{A^2}{4} T_b$$

The receiver decodes the kth transmitted bit as 1 if the output at the kth signaling interval is greater than T_0^*, and as 0 otherwise.

The probability of error P_e can be computed using Equations (8.20) and (8.21) as

$$\gamma_{\max}^2 = \int_{-\infty}^{\infty} \frac{|P(f)|^2}{G_n(f)}\, df$$
$$= \frac{2}{\eta} \int_0^{T_b} p^2(t)\, dt$$
$$= \frac{2}{\eta} \int_0^{T_b} A^2 \cos^2 \omega_c t\, dt$$
$$= \frac{A^2 T_b}{\eta}$$

and

$$P_e = \int_{\frac{1}{2}\gamma_{\max}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz = Q\left(\sqrt{\frac{A^2 T_b}{4\eta}}\right) \tag{8.31}$$

The signal $s_2(t)$ is present at the receiver input only one half the time on the average, and for the remaining half there is no signal since $s_1(t) = 0$. Hence the average signal power at the receiver input is given by

$$S_{av} = A^2/4$$

We can express the probability of error in terms of the average signal power as

$$P_e = Q\left(\sqrt{\frac{S_{av} T_b}{\eta}}\right) \tag{8.32}$$

The probability of error is sometimes expressed in terms of the average signal energy per bit, $E_{av} = (S_{av})T_b$, as

$$P_e = Q(\sqrt{E_{av}/\eta}) \tag{8.33}$$

Example 8.2. Binary data has to be transmitted over a telephone link that has a usable bandwidth of 3000 Hz and a maximum achievable signal-to-noise power ratio of 6 dB at its output. (a) Determine the maximum signaling rate and P_e if a coherent ASK scheme is used for transmitting binary data through this channel. (b) If the data rate is maintained at 300 bits/sec, calculate the error probability.

Solution

(a) If we assume that an ASK signal requires a bandwidth of $3r_b$ Hz, then the maximum signaling rate permissible is $r_b = 1000$ bits/sec. The probability of error can be computed as follows:

Average signal power $= A^2/4$

Noise power $= (2)(\eta/2)(3000)$

$$\frac{\text{Average signal power}}{\text{Noise power}} = 4 = \frac{A^2}{12{,}000\eta} \quad \text{or} \quad \frac{A^2}{\eta} = 48{,}000$$

Hence, $A^2/4\eta r_b = 12$ and

$$P_e = Q(\sqrt{12}) = Q(3.464) \simeq 0.0003$$

(b) If the bit rate is reduced to 300 bits/sec, then

$$\frac{A^2}{4\eta r_b} = 40$$

and

$$P_e = Q(\sqrt{40}) = Q(6.326) \simeq 10^{-10}.$$

8.3.2 Noncoherent ASK

In ideal coherent detection of ASK signals, we assume that there is available at the receiver an exact replica of the arriving signal. That is, we have assumed that a phase coherent local carrier can be generated at the receiver. While this may be possible by the use of very stable oscillators in both the transmitter and receiver, the cost may be excessive.

Noncoherent detection schemes do not require a phase-coherent local oscillator signal. These schemes involve some form of rectification and lowpass filtering at the receiver. The block diagram of a noncoherent receiver for the ASK signaling scheme is shown in Figure 8.9. The computation of the error probability for this receiver is more difficult because of the nonlinear operations that take place in the receiver. In the following analysis we will rely heavily on the results derived in Chapter 3. The input to the receiver is

$$V(t) = \begin{cases} A\cos\omega_c t + n_i(t) & \text{when } b_k = 1 \\ n_i(t) & \text{when } b_k = 0 \end{cases}$$

where $n_i(t)$ is the noise at the receiver input, which is assumed to be zero mean, Gaussian, and white. Now, if we assume the bandpass filter to have a bandwidth of $2/T_b$ centered at f_c, then it passes the signal component without much distortion. At the filter output we have

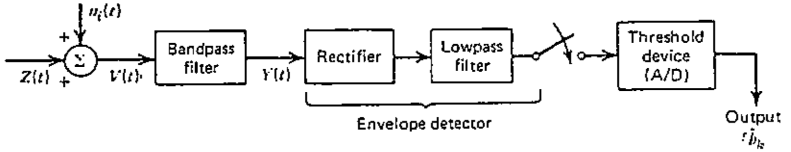


Figure 8.9 Noncoherent ASK receiver.

$$\begin{aligned} Y(t) &= A_k\cos\omega_c t + n(t) \\ &= A_k\cos\omega_c t + n_c(t)\cos\omega_c t - n_s(t)\sin\omega_c t \end{aligned} \tag{8.34}$$

where $A_k = A$ when the kth transmitted bit $b_k = 1$ and $A_k = 0$ when $b_k = 0$. $n(t)$ is the noise at the output of the bandpass filter and $n_c(t)$, $n_s(t)$ are the quadrature components of the narrowband noise $n(t)$ (Chapter 3). We can rewrite Equation (8.34) in envelope and phase form as

$$Y(t) = R(t)\cos[\omega_c t + \theta(t)]$$

where

$$R(t) = \sqrt{[A_k + n_c(t)]^2 + [n_s(t)]^2} \tag{8.35}$$

Assuming ideal operation, the output of the envelope detector is $R(t)$ and the transmitted bit sequence $\{b_k\}$ is recovered from $R(kT_b)$.

To calculate the probability of error, we need to determine the conditional probability density functions $f_{R|b_k=0}(r)$ and $f_{R|b_k=1}(r)$, and the optimum value of the threshold. Using the results derived in Chapter 3, we obtain these conditional pdfs as

$$f_{R|b_k=0}(r) = \frac{r}{N_0}\exp\left(-\frac{r^2}{2N_0}\right), \quad r > 0 \tag{8.36a}$$

$$f_{R|b_k=1}(r) = \frac{r}{N_0}I_0\left(\frac{Ar}{N_0}\right)\exp\left(-\frac{r^2 + A^2}{2N_0}\right), \quad r > 0 \tag{8.36b}$$

where N_0 is the noise power at the output of the bandpass filter

$$N_0 = \eta B_T \approx 2\eta/T_b$$

and $I_0(x)$ is the modified Bessel function of the first kind and zero order defined by

$$I_0(x) = \frac{1}{2\pi}\int_0^{2\pi} \exp(x\cos(u))\,du$$

In order for the envelope detector to operate above the noise threshold

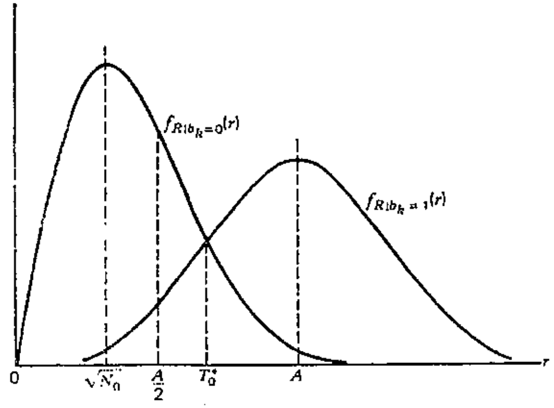


Figure 8.10 Pdf's of the envelope of the noise and the envelope of the signal plus noise.

(Chapter 6, Section 6.3), the carrier amplitude at the receiver input should be such that $A^2 \gg N_0$. If we assume $A^2 \gg N_0$, then we can approximate the Bessel function by

$$I_0\left(\frac{Ar}{N_0}\right) \approx \sqrt{\frac{N_0}{2\pi Ar}}\exp\left(\frac{Ar}{N_0}\right)$$

Hence,

$$f_{R|b_k=1}(r) = \sqrt{\frac{r}{2\pi AN_0}}\exp\left(-\frac{(r-A)^2}{2N_0}\right), \quad r > 0$$

which is essentially a Gaussian distribution with mean A and variance N_0 since $r/2\pi AN_0 \approx 1/2\pi N_0$ in the vicinity of $r = A$, where the pdf has the bulk of its area. Sketches of the pdfs are given in Figure 8.10.

The receiver compares the output of the envelope detector $R(t)$ with a threshold value T_0 and decodes the received signal as $s_1(t)$ if $R(kT_b) \leq T_0$ and as $s_2(t)$ if $R(kT_b) > T_0$. Clearly, the threshold T_0 should be set between 0 and A such that the probability of error is minimized. It can be shown that the value of T_0, say T_0^*, which minimizes the probability of error has to satisfy

$$f_{R|b_k=0}(T_0^*) = f_{R|b_k=1}(T_0^*) \tag{8.37a}$$

The relationship

$$T_0^* \approx \frac{A}{2}\sqrt{1+\frac{8N_0}{A^2}} \tag{8.37b}$$

is an excellent analytic approximation to the solution of Equation (8.37*a*). When the carrier amplitude at the receiver input is such that $A^2 \gg N_0$, then $T_0^* \approx A/2$.

The probability of error P_e is given by

$$P_e = \tfrac{1}{2}P(\text{error}|b_k = 0) + \tfrac{1}{2}P(\text{error}|b_k = 1)$$
$$= \tfrac{1}{2}P_{e0} + \tfrac{1}{2}P_{e1}$$

where

$$P_{e0} = \int_{A/2}^{\infty} \frac{r}{N_0}\exp\left(-\frac{r^2}{2N_0}\right)dr = \exp\left(-\frac{A^2}{8N_0}\right) \tag{8.38a}$$

and

$$P_{e1} \approx \int_{-\infty}^{A/2} \frac{1}{\sqrt{2\pi N_0}}\exp\left(-\frac{(r-A)^2}{2N_0}\right)dr$$
$$= Q\left(\frac{A}{2\sqrt{N_0}}\right) \tag{8.38b}$$

Using the approximation

$$Q(x) = \frac{\exp(-x^2/2)}{x\sqrt{2\pi}}$$

for large x, we can reduce P_{e1} to the form

$$P_{e1} \approx \sqrt{\frac{4N_0}{2\pi A^2}}\exp\left(-\frac{A^2}{8N_0}\right) \tag{8.39a}$$

Hence,

$$P_e \approx \tfrac{1}{2}\left[1+\sqrt{\frac{4N_0}{2\pi A^2}}\right]\exp\left(-\frac{A^2}{8N_0}\right)$$
$$\approx \tfrac{1}{2}\exp\left(-\frac{A^2}{8N_0}\right) \quad \text{if } A^2 \gg N_0 \tag{8.39b}$$

where $N_0 = \eta B_T$ and B_T is the bandwidth of the bandpass filter.

The probability of error for the noncoherent ASK receiver will always be higher than the error probability of a coherent receiver operating with the same signal power, signaling speed, and noise psd. However, the noncoherent receiver is much simpler than the coherent receiver.

In order to obtain optimum performance, the threshold value at the receiver should be adjusted according to Equation (8.37*b*) as the signal level changes. Furthermore, the filters used in the receiver should be discharged, via auxiliary circuitry, at the end of each bit interval in order to reduce inter symbol interference. While the resulting circuit is no longer a linear time-invariant filter, it does act like a liner time-invariant filter in between discharge intervals. For such a rapidly discharging filter, the filter bandwidth is no longer critical with respect to intersymbol interference.

In the noncoherent receiver shown in Figure 8.9, we have assumed that timing information is available at the receiver. This timing information is usually extracted from the envelope of the received signal using a technique similar to the one described in Chapter 5, Section 5.7.

It is worth noting here that in the noncoherent ASK scheme, the probability of incorrectly decoding "1" is different from the probability of incorrectly decoding "0." Thus the noncoherent ASK scheme results in a nonsymmetric binary channel (Chapter 4, Section 4.5).

Example 8.3. Binary data is transmitted over an RF bandpass channel with a usable bandwidth of 10 MHz at a rate of (4.8) (10^6) bits/sec using an ASK signaling method. The carrier amplitude at the receiver antenna is 1 mv and the noise power spectral density at the receiver input is 10^{-15} watt/Hz. (a) Find the error probability of a coherent receiver, and (b) find the error probability of a noncoherent receiver.

Solution

(a) The bit error probability for the coherent demodulator is

$$P_e = Q\left(\sqrt{\frac{A^2T_b}{4\eta}}\right); \qquad A = 1\,\text{mv}, \quad T_b = 10^{-6}/4.8$$

$$\eta/2 = 10^{-15}\ \text{watt/Hz}$$

Hence, $P_e = Q(\sqrt{26}) \simeq 2(10^{-7})$.

(b) The noise power at the filter output is $N_0 = 2\eta r_b = 1.92(10^{-8})$ and $A^2 = 10^{-6}$. Hence, $A^2 \gg N_0$ and we can use the approximations given in the preceding paragraphs for P_{e0} and P_{e1}

$$P_{e1} = Q\left(\sqrt{\frac{A^2}{4N_0}}\right) = Q(3.61) = 0.0002$$

$$P_{e0} = \exp\left(-\frac{A^2}{8N_0}\right) \approx 0.0015$$

Hence, $P_e = \frac{1}{2}(P_{e0} + P_{e1}) = 0.00085$.

8.4 BINARY PSK SIGNALING SCHEMES

Phase-shift keying, or discrete phase modulation, is another technique available for communicating digital information over bandpass channels. In PSK

signaling schemes the waveforms $s_1(t) = -A\cos\omega_c t$ and $s_2(t) = A\cos\omega_c t$ are used to convey binary digits 0 and 1, respectively. The binary PSK waveform $Z(t)$ can be described by

$$Z(t) = D(t)(A\cos\omega_c t)$$

where $D(t)$ is a random binary waveform with period T_b and levels -1 and 1. The only difference between the ASK and PSK waveform is that in the ASK scheme the carrier is switched on and off whereas in the PSK scheme the carrier is switched between levels $+A$ and $-A$. The power spectral density of the PSK signal can be shown to be

$$G_Z(f) = \frac{A^2}{4}[G_D(f-f_c) + G_D(f+f_c)]$$

where

$$G_D(f) = \frac{\sin^2 \pi f T_b}{\pi^2 f^2 T_b} \tag{8.40}$$

Comparison of Equation (8.40) with Equation (8.30) reveals that the shapes of the psd of the binary PSK signal and the ASK signal are similar. The only difference is that the PSK spectrum does not have an impulse at the carrier frequency. The bandwidth requirement of the PSK signal is the same as that of the ASK signal. The similarity between the ASK and PSK is somewhat misleading. The ASK is a linear modulation scheme whereas the PSK, in the general case, is a nonlinear modulation scheme.

The primary advantage of the PSK signaling scheme lies in its superior performance over the ASK scheme operating under the same peak power limitations and noise environment. In the following sections we will derive expressions for the probability of error for coherent and noncoherent PSK signaling schemes.

8.4.1 Coherent PSK

The transmitted bit sequence $\{b_k\}$ can be recovered from the PSK signal using the integrate and dump correlation receiver shown in Figure 8.7 with a local reference signal $s_2(t) - s_1(t) = 2A\cos\omega_c t$ that is synchronized in phase and frequency with the incoming signal. The signal components of the receiver output at $t = kT_b$ are

$$s_{01}(kT_b) = \int_{(k-1)T_b}^{kT_b} s_1(t)[s_2(t) - s_1(t)]\,dt = -A^2T_b$$

$$s_{02}(kT_b) = \int_{(k-1)T_b}^{kT_b} s_2(t)[s_2(t) - s_1(t)]\,dt = A^2T_b$$

The optimum threshold setting is $T_0^* = 0$, which is independent of the carrier strength at the receiver input. The probability of error P_e is given by

$$P_e = Q(\gamma_{\max}/2)$$

where

$$\gamma_{\max}^2 = \frac{2}{\eta}\int_0^{T_b} (2A\cos\omega_c t)^2\, dt = \frac{4A^2T_b}{\eta}$$

or

$$P_e = Q(\sqrt{A^2T_b/\eta}) \tag{8.41}$$

The average signal power S_{av} and the signal energy per bit E_{av} for the PSK scheme are

$$S_{av} = A^2/2$$

and

$$E_{av} = (A^2/2)T_b$$

We can express the probability of error in terms of S_{av} and E_{av} as

$$P_e = (\sqrt{2S_{av}T_b/\eta}) \tag{8.42}$$

$$= Q(\sqrt{2E_{av}/\eta}) \tag{8.43}$$

Comparing the probability of error for the coherent PSK (Equation (8.42)) with the probability of error for the coherent ASK (Equation (8.32)), we see that for equal probability of error the average signal power of the ASK signal should be twice the average power of the PSK signal. That is, the coherent PSK scheme has a 3-dB power advantage over the coherent ASK scheme.

8.4.2 Differentially Coherent PSK

The differentially coherent PSK (DPSK) signaling scheme makes use of a clever technique designed to get around the need for a coherent reference signal at the receiver. In the DPSK scheme, the phase reference for demodulation is derived from the phase of the carrier during the preceding signaling interval, and the receiver decodes the digital information based on the differential phase. If the channel perturbations and other disturbances are slowly varying compared to the bit rate, then the phase of the RF pulses $s(t)$ and $s(t-T_b)$ are affected by the same manner, thus preserving the information contained in the phase difference. If the digital information had been *differentially encoded* in the carrier phase at the transmitter, the decoding at the receiver can be accomplished without a coherent local oscillator signal. The DPSK scheme may be thought of as the noncoherent version of the PSK scheme discussed in the preceding section.

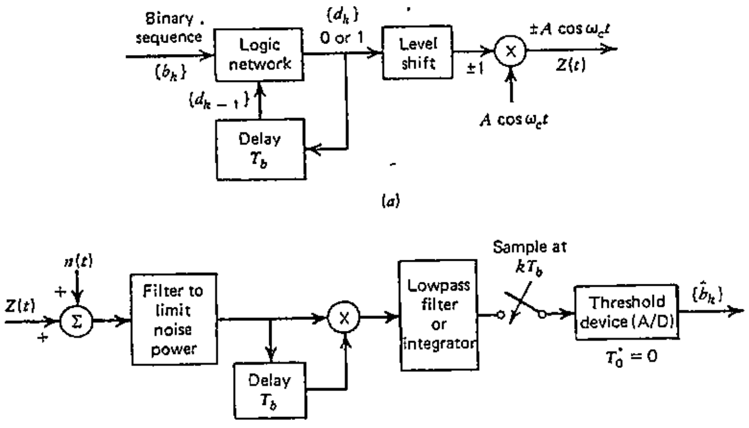


Figure 8.11 (a) DPSK modulator. (b) DPSK demodulator.

Block diagrams of a DPSK modulator and demodulator are shown in Figures 8.11*a* and 8.11*b*, respectively. The differential encoding operation performed by the modulator is explained in Table 8.2. The encoding process starts with an arbitrary first bit, say 1, and thereafter the encoded bit stream d_k is generated by

$$d_k = d_{k-1}b_k \oplus \bar{d}_{k-1}\bar{b}_k$$

Table 8.2. Differential encoding and decoding

Input sequence (b_k)		1	1	0	1	0	0	0	1	1
Encoded sequence (d_k)	1[a]	1	1	0	0	1	0	1	1	1
Transmitted phase	0	0	0	π	π	0	π	0	0	0
Phase comparison output		+	+	−	+	−	−	−	+	+
Output bit sequence		1	1	0	1	0	0	0	1	1

[a] Arbitrary starting reference bit.

The differential sequence d_k then phase-shift keys a carrier with the phases 0 and π, as shown in Table 8.2.

The DPSK receiver correlates the received signal plus noise with a delayed version (delay = 1-bit duration) of the signal plus noise. The output of the correlator is compared with zero and a decision is made in favor of 1 or 0 depending on whether the correlator output is + or −, respectively. The reader can easily verify that the receiver recovers the bit sequence $\{b_k\}$ correctly, in the absence of noise, by assuring himself that the receiver essentially checks to see if the phase angles of the received carrier during two successive bit intervals are the same or different. With an initial angle of 0 (for the reference bit), the receiver output is 1 at the end of the kth signaling interval if the carrier phase is the same during the $(k-1)$st and the kth signaling intervals. If the phase angles are different, then the receiver output is 0. The last two rows in Table 8.2 illustrate that phase comparison detection at the receiver works correctly.

The noise performance of the DPSK might appear to be inferior compared to coherent PSK because the phase reference is contaminated by noise in the DPSK scheme. However, the perturbations in phase reference due to noise tend to cancel out and the degradation in performance is not too great. In the following paragraphs we will derive an expression for the probability of error for the DPSK scheme.

For purposes of analysis, let us assume that the carrier phase during the $(k-1)$st and the kth signaling intervals is 0, that is, $\phi_{k-1} = \phi_k = 0$. An error in decoding the kth bit occurs if the phase comparator output is negative. The input to the lowpass filter in Figure 8.11b can be written as

$$q(t) = [A\cos\omega_c t + n_f(t)][A\cos\omega_c t' + n_f(t')], \quad (k-1)T_b \leqslant t \leqslant kT_b$$

where $t' = t - T_b$, and $n_f(t)$ is the response of the front-end filter to $n(t)$. Substituting the quadrature representation

$$n_f(t) = n_c(t)\cos\omega_c t - n_s(t)\sin\omega_c t$$

in the preceding equation, we obtain

$$\begin{aligned} q(t) = {} & [A + n_c(t)]\cos\omega_c t[A + n_c(t')]\cos\omega_c t' \\ & - [A + n_c(t)]\cos\omega_c t[n_s(t')\sin\omega_c t'] \\ & - n_s(t)\sin\omega_c t[A + n_c(t')\cos\omega_c t'] \\ & + n_s(t)n_s(t')\sin\omega_c t\sin\omega_c t' \end{aligned}$$

The reader can verify that the lowpass filter output $V_0(kT_b)$ is given by [remember that $\omega_c T_b = k\pi (k \geqslant 2)$, hence $\sin\omega_c t = \sin\omega_c t'$ and $\cos\omega_c t = \cos\omega_c t'$]

$$V_0(kT_b) = c[A + n_c(t)][A + n_c(t')] + n_s(t)n_s(t')$$

where c is a positive constant $t = kT_b$, and $t' = (k-1)T_b$. The probability of error P_e is given by

$$P_e = P[V_0(kT_b) < 0] = P\left(\frac{1}{c}V_0(kT_b) < 0\right)$$

In order to simplify the expression for the probability of error, let us define

$$\alpha = A + \frac{n_c(t) + n_c(t')}{2}, \qquad \beta = \frac{n_c(t) - n_c(t')}{2}$$

$$\nu = \frac{n_s(t) + n_s(t')}{2}, \qquad \delta = \frac{n_s(t) - n_s(t')}{2}$$

We now have

$$\frac{1}{c}V_0(kT_b) = (\alpha^2 + \beta^2) - (\nu^2 + \delta^2)$$

$$\begin{aligned} P\left(\frac{1}{c}V_0(kT_b) < 0\right) &= P(\alpha^2 + \beta^2 < \nu^2 + \delta^2) \\ &= P(\sqrt{\alpha^2 + \beta^2} < \sqrt{\nu^2 + \delta^2}) \end{aligned}$$

If we denote $\sqrt{\alpha^2 + \beta^2}$ by X_1 and $\sqrt{\nu^2 + \delta^2}$ by X_2, then X_1 has a Rice pdf and X_2 has a Raleigh pdf, and the probability of error P_e has the form

$$P_e = P(X_1 < X_2) = \int_0^\infty P(X_2 > x_1 | X_1 = x_1) f_{X_1}(x_1)\,dx_1$$

where

$$P(X_2 > x_1 | X_1 = x_1) = \int_{x_1}^\infty f_{X_2}(x_2)\,dx_2$$

since X_1 and X_2 are statistically independent. The pdf's involved in the preceding expressions have the forms given in Equations (8.36a) and (8.36b) and the probability of error can be shown to be equal to (see Problem 8.17)

$$P_e = \tfrac{1}{2}\exp(-A^2T_b/2\eta) \tag{8.44}$$

The example given below shows that the DPSK scheme requires 1 dB more power than the coherent PSK scheme when the error probabilities of both systems are of the order of 10^{-4}. The slight increase in power requirements for the DPSK signaling method is more than offset by the fact that DPSK does not require a coherent reference signal at the receiver. Because of the fixed delay in the DPSK receiver, the system is locked on a specific signaling speed, thus precluding asynchronous data transmission. Another minor problem in DPSK schemes is that errors tend to propagate, at least to adjacent bits, due to the correlation between signaling waveforms and the noise over adjacent signaling intervals.

Example 8.4. Binary data is transmitted at a rate of 10^6 bits/sec over a microwave link having a bandwidth of 3 MHz. Assume that the noise power spectral density at the receiver input is $\eta/2 = 10^{-10}$ watt/Hz. Find the average carrier power required at the receiver input for coherent PSK and DPSK signaling schemes to maintain $P_e \leqslant 10^{-4}$.

Solution
The probability of error for the PSK scheme is

$$(P_e)_{\text{PSK}} = Q(\sqrt{2S_{av}T_b/\eta}) \leqslant 10^{-4},$$

This requires

$$\sqrt{2S_{av}T_b/\eta} \geqslant 3.75$$

or

$$(S_{av})_{\text{PSK}} \geqslant (3.75)^2(10^{-10})(10^6) = 1.48\ \text{dBm}$$

For the DPSK scheme we have

$$(P_e)_{\text{DPSK}} = \tfrac{1}{2}\exp[-(A^2T_b/2\eta)] \leqslant 10^{-4},$$

hence

$$S_{av}T_b/\eta \geqslant 8.517 \quad \text{or} \quad (S_{av})_{\text{DPSK}} \geqslant 2.313\ \text{dBm}$$

This example illustrates that the DPSK signaling scheme requires about 1 dB more power than the coherent PSK scheme when the error probability is of the order of 10^{-4}.

8.5 BINARY FSK SIGNALING SCHEMES

FSK signaling schemes find a wide range of applications in low-speed digital data transmission systems. Their appeal is mainly due to hardware advantages that result principally from the use of a noncoherent demodulation process and the relative ease of signal generation. As we will see later in this section, FSK schemes are not as efficient as PSK schemes in terms of power and bandwidth utilization. In the binary FSK signaling scheme, the waveforms $s_1(t) = A\cos(\omega_c t - \omega_d t)$ and $s_2(t) = A\cos(\omega_c t + \omega_d t)$ are used to convey binary digits 0 and 1, respectively. The information in an FSK signal is essentially in the frequency of the signal.

The binary FSK waveform is a continuous phase constant envelope FM waveform. The binary FSK waveform can be mathematically represented as follows:

$$Z(t) = A\cos\left(\omega_c t + \omega_d \int_{-\infty}^{t} D(t')\,dt' + \theta\right) \tag{8.45}$$

where $D(t)$ is a random binary waveform with levels +1 when $b_k = 1$ and −1 when $b_k = 0$, and θ is the phase angle of the carrier at time $t = 0$. The instantaneous frequency of the binary FSK signal is given by

$$\begin{aligned} f_i &= \frac{d}{dt}[\text{phase of } Z(t)] \\ &= \omega_c + \omega_d D(t) \end{aligned}$$

Since $D(t) = \pm 1$, the instantaneous frequency ω_i has two values: $\omega_i = \omega_c \pm \omega_d$.

The derivation of the power spectral density of the digital FM waveform is rather involved and hence we will look only at the results of the derivation shown in Figure 8.12 (see Lucky's book, Chapter 8, for a detailed derivation).

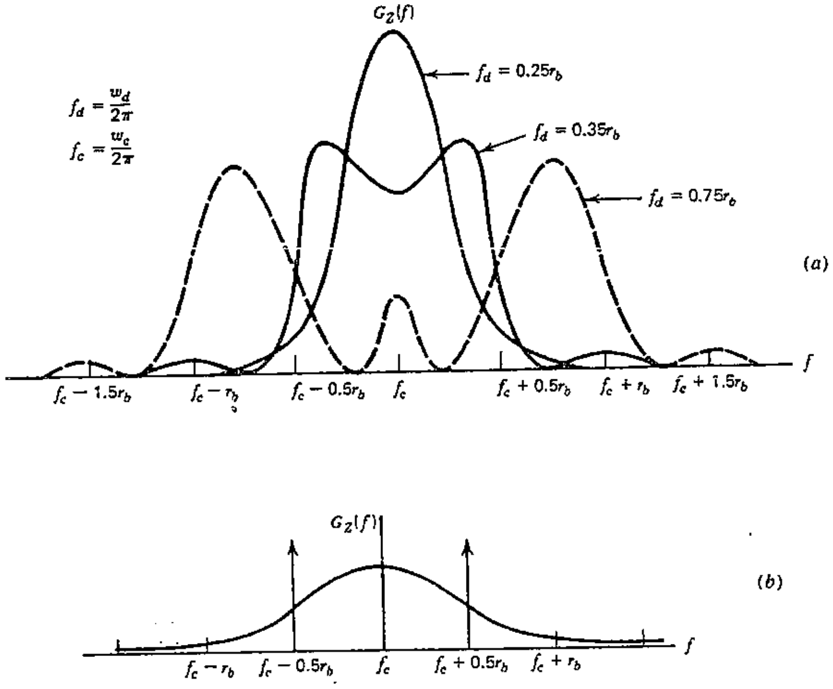


Figure 8.12 (a) Power spectral density of FSK signals. (b) Power spectral density of a binary FSK signal with $2f_d = r_b$.

The power spectral density curves displayed in Figure 8.12*a* exhibit the following characteristics: For low values of f_d/r_b the curve has a smooth roll off with a peak at the carrier frequency. The FSK signal bandwidth in this case is of the order of $2r_b$ Hz, which is the same as the order of bandwidth of the PSK signal. As f_d/r_b increases, major peaks in the power spectral density curve occur at the transmit frequencies f_c+f_d and f_c-f_d and the bandwidth of the signal exceeds $2r_b$, the bandwidth of the PSK signal. For large values of f_d/r_b, the FSK signal essentially consists of two interleaved ASK signals of differing carrier frequencies, say f_c+f_d and f_c-f_d. Further, when $2f_d=mr_b$, m an integer, the psd has impulses corresponding to discrete frequency sinusoidal components as shown in Figure 8.12*b*. In general, we can say that the bandwidth of the FSK signal is greater than the bandwidth of the ASK and the PSK signals.

As mentioned earlier, the binary FSK waveform given in Equation (8.45) is a continuous phase waveform. In order to maintain phase continuity, the phase at every transition is made to be dependent on the past data sequence. To visualize how one could generate a continuous phase constant envelope FSK signal, consider the following waveform construction procedure: The sequence $\{b_k\}$ is used to generate a sequence of segmented cosine waveforms $A\cos(\omega_c t+\omega_k t+\theta_k)$, where $\omega_k=+\omega_d$ if $b_k=1$ and $\omega_k=-\omega_d$ if $b_k=0$. The FM waveform given in Equation (8.45) is then constructed by specifying the sequence $\{\theta_k\}$ as follows: Let θ_1 be arbitrarily set equal to some value θ. Then $\theta_2=\theta+(\omega_1+\omega_c)T_b$, $\theta_3=\theta+(\omega_1+\omega_c)T_b+(\omega_2+\omega_c)T_b,\ldots,$ and $\theta_n=\theta+(\omega_1+\omega_c)T_b+\cdots+(\omega_{n-1}+\omega_c)T_b$. By shifting the phase of the different segments, one obtains a continuous phase constant envelope FM wave.

It is also possible to generate a digital FM wave ignoring the phase continuity. This can be done by having two oscillators tuned to $\omega_c+\omega_d$ and $\omega_c-\omega_d$ whose outputs are directly controlled by the digital baseband waveform $D(t)$. This method gives rise to undesirable transients in addition to complicating the transmitter. For these reasons, this method is rarely used in practice.

The FSK signal $Z(t)$ can be demodulated using a coherent correlation receiver or using a suboptimal scheme consisting of bandpass filters and envelope detectors. Correlation detection of FSK signals is very seldom used; our study of coherent FSK is mainly for comparison purposes.

8.5.1 Coherent FSK

If the FSK signal is demodulated using the correlation receiver shown in Figure 8.7, the local carrier signal required is

$$s_2(t)-s_1(t)=A\cos(\omega_c t+\omega_d t)-A\cos(\omega_c t-\omega_d t)$$

The input to the A/D converter at sampling time $t=kT_b$ is $s_{01}(kT_b)$ or $s_{02}(kT_b)$, where

$$s_{02}(kT_b)=\int_0^{T_b}s_2(t)[s_2(t)-s_1(t)]\,dt$$

$$s_{01}(kT_b)=\int_0^{T_b}s_1(t)[s_2(t)-s_1(t)]\,dt$$

If the signal energy E_1 and E_2 are the same, then $s_{02}(kT_b)=-s_{01}(kT_b)$ and hence the receiver threshold setting is at 0. The probability of error P_e for the correlation receiver is given by (from Equation (8.21))

$$P_e=Q(\gamma_{\max}/2)$$

where

$$\gamma_{\max}^2=\frac{2}{\eta}\int_0^{T_b}[s_2(t)-s_1(t)]^2\,dt$$

Substituting $s_2(t)=A\cos(\omega_c t+\omega_d t)$ and $s_1(t)=A\cos(\omega_c t-\omega_d t)$ and performing the integration, we get

$$\gamma_{\max}^2=\frac{2A^2T_b}{\eta}\left(1-\frac{\sin 2\omega_dT_b}{2\omega_dT_b}+\frac{1}{2}\frac{\sin[2(\omega_c+\omega_d)T_b]}{2(\omega_c+\omega_d)T_b}\right.$$
$$\left.-\frac{1}{2}\frac{\sin[2(\omega_c-\omega_d)T_b]}{2(\omega_c-\omega_d)T_b}-\frac{\sin 2\omega_cT_b}{2\omega_cT_b}\right) \qquad (8.46a)$$

If we make the following assumptions:

$$\omega_cT_b\gg 1, \qquad \omega_c\gg\omega_d$$

which are usually encountered in practical systems, then the last three terms in Equation (8.46a) may be neglected. We now have

$$\gamma_{\max}^2=\frac{2A^2T_b}{\eta}\left(1-\frac{\sin 2\omega_dT_b}{2\omega_dT_b}\right) \qquad (8.46b)$$

The quantity $\gamma_{\max}^2$ in the preceding equation attains the largest value when the frequency offset ω_d is selected so that $2\omega_dT_b=3\pi/2$. For this value of ω_d we find

$$\gamma_{\max}^2=(2.42)(A^2T_b/\eta)$$

and

$$P_e=Q(\sqrt{0.61(A^2T_b/\eta)}) \qquad (8.47)$$

Once again, if we define $S_{av}=A^2/2$ and $E_{av}=A^2T_b/2$, then we can express P_e as

$$P_e=Q(\sqrt{1.2S_{av}T_b/\eta})$$
$$=Q(\sqrt{1.2E_{av}/\eta}) \qquad (8.48)$$

Comparison of the probability of error for the coherent FSK scheme with the error probability for coherent PSK scheme (Equation (8.42)) shows that coherent FSK requires about 2.2 dB more power than the coherent PSK scheme. The FSK signal also uses more bandwidth than the PSK signal. Thus coherent FSK does not have any advantages over the coherent PSK scheme.

8.5.2 Noncoherent FSK

Since the FSK scheme can be thought of as the transmission of two interleaved ASK signals (assuming that $2f_d = mr_b$, m an integer), the first with a carrier frequency $f_c - f_d$ and the second with carrier frequency $f_c + f_d$, it should be possible to detect the signal using two bandpass filters with center frequencies $f_c + f_d$ and $f_c - f_d$. Such a detection scheme is shown in Figure 8.13. The probability of error for the noncoherent FSK receiver can be derived easily using the results derived in Section 8.3.2 for the noncoherent ASK receiver. As a matter of fact, the derivation is somewhat simpler since we do not have to face the problem of calculating the optimum threshold setting. Because of symmetries, the threshold is set at zero in noncoherent FSK receivers.

Assuming that $s_1(t) = A\cos(\omega_c - \omega_d)t$ has been transmitted during the kth signaling interval, the pdf of the envelope $R_1(kT_b)$ of the bottom filter is

$$f_{R_1|s_1(t)}(r_1) = \frac{r_1}{N_0} I_0\left(\frac{Ar_1}{N_0}\right) \exp\left(-\frac{r_1^2 + A^2}{2N_0}\right), \quad r_1 > 0$$

where $N_0 = \eta B_T$, and B_T is the filter bandwidth. The top filter responds to noise alone and therefore $R_2(kT_b)$ has a Rayleigh pdf given by

$$f_{R_2|s_1(t)}(r_2) = \frac{r_2}{N_0} \exp\left(\frac{-r_2^2}{2N_0}\right), \quad r_2 > 0$$

An error occurs when $R_2 > R_1$, and this error probability is obtained as

$$\begin{aligned} P\,[\text{error}|s_1(t)\text{sent}] &= P(R_2 > R_1) \\ &= \int_0^\infty f_{R_1|s_1}(r_1)\left[\int_{r_1}^\infty f_{R_2|s_1}(r_2)\,dr_2\right] dr_1 \end{aligned} \qquad (8.49)$$

since the random variables $R_1(kT_b)$ and $R_2(kT_b)$ will be independent if $f_d = mr_b/4$, where m is an integer (Problem 8.24). By symmetry, we have $P\,[\text{error}|s_1(t)\text{ sent}] = P\,[\text{error}|s_2(t)\text{ sent}]$ so that

$$P(\text{error}) = P_e = P\,[\text{error}|s_1(t)\text{ sent}]$$

Substituting the appropriate pdf's in Equation (8.49) and carrying out the integration (with the help of a table of definite integrals), we obtain

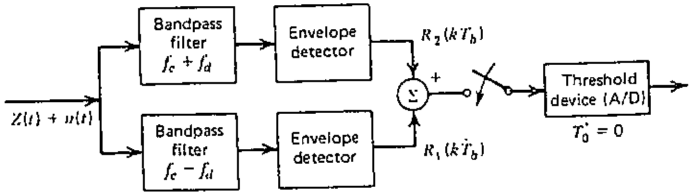


Figure 8.13 Noncoherent demodulation of binary FSK signals.

$$P_e = \tfrac{1}{2}\exp(-A^2/4N_0) \qquad (8.50)$$

The filter bandwidth is usually of the order of $2/T_b$ and hence N_0 in Equation (8.50) is approximately equal to $2\eta/T_b$.

The error probability for a noncoherent FSK receiver will be higher than the error probability of a coherent FSK receiver. However, because of its simplicity, the noncoherent FSK scheme is widely used in practice.

Example 8.5. Binary data is transmitted over a telephone line with usable bandwidth of 2400 Hz using the FSK signaling scheme. The transmit frequencies are 2025 and 2225 Hz, and the data rate is 300 bits/sec. The average signal-to-noise power ratio at the output of the channel is 6 dB. Calculate P_e for the coherent and noncoherent demodulation schemes.

Solution. We are given $r_b = 300$, $f_c + f_d = 2225$ Hz, and $f_c - f_d = 2025$ Hz. Hence, $f_c = 2125$ Hz and $f_d = 100$ Hz. Before we use Equation (8.48) to obtain P_e, we need to make sure that the assumptions $\omega_c T_b \gg 1$, $\omega_c \gg \omega_d$, and $2\omega_d T_b \approx 3\pi/2$ are valid. The reader can verify that all these assumptions are satisfied. Now, we are given that $S/N = (A^2/2)/(2400\eta) = 4$ or $A^2T_b/\eta = 64$. Using Equation (8.48), we obtain

$$(P_e)_{\substack{\text{coh}\\\text{FSK}}} = Q(\sqrt{(0.61)64}) \approx 10^{-9}$$

For the noncoherent scheme, P_e is given by Equation (8.50) as

$$(P_e)_{\substack{\text{noncoh}\\\text{FSK}}} = \tfrac{1}{2}\exp\left(-\frac{A^2T_b}{8\eta}\right) = \frac{e^{-8}}{2} = 1.68(10^{-4})$$

8.6 COMPARISON OF DIGITAL MODULATION SYSTEMS

We have developed formulas in the preceding sections that relate the performance of various modulation schemes, as measured by the probability of error, to parameters of the system, such as signaling rate, noise power spectral density, and signal power. We also discussed the complexity of equipment required to generate, transmit, and demodulate the different types of modulated signals. In this section we will attempt to compare the performance of various digital modulation schemes.

We begin our comparison by emphasizing that the choice of a modulation method depends on the specific application. The choice may be based on relative immunity to noise and channel impairments (such as nonlinearities, phase jitter, fading, and frequency offset), simplicity of equipment, and compatibility with other equipment already in place in the system. While it is not our intent to compare modulation systems under all conditions cited above, we will however attempt to provide the reader with some guidelines that might be useful in comparing and selecting a modulation scheme for a particular application. We will compare systems operating at the same data rate (r_b), probability of error (P_e), and noise environment.

8.6.1 Bandwidth Requirements

If one is interested in high speed data transmission over a noisy bandpass channel, then vestigial-sideband (VSB) modulation with baseband signal shaping is a better choice than ASK, PSK, or FSK schemes for efficient bandwidth utilization. Bandwidth requirements of VSB schemes with baseband signal shaping are of the order of r_b. The bandwidth requirements of ASK and PSK schemes are of the order of $2r_b$, whereas the bandwidth of the FSK signal is somewhat larger than $2r_b$. Thus if bandwidth is of primary concern, the FSK scheme is generally not considered.

8.6.2 Power Requirements

The power requirements of various schemes can be compared using the relationships derived in the preceding sections. These relationships are summarized in Table 8.3 and plots of the probability of error P_e versus $A^2T_b/2\eta$ are shown in Figure 8.14. The horizontal axis in Figure 8.14 should be read as peak received (or transmitted) power and the peak power A^2 is the same for all schemes. The error probability in most practical systems is in the range of 10^{-4} to 10^{-7} and hence we will do our comparison of power requirements assuming that $10^{-7} < P_e < 10^{-4}$.

Table 8.3. Comparison of Binary digital modulation schemes

Scheme	$s_1(t), s_2(t)$	BW	P_e	S/N for $P_e = 10^{-4}$ (dB)	Equipment complexity	Comments
Coherent ASK	$s_1(t) = A\cos\omega_c t$ $s_2(t) = 0$ $\omega_c = k2\pi r_b$ k-integer	$\simeq 2r_b$	$Q\left(\sqrt{\frac{A^2T_b}{4\eta}}\right)$	14.45	Moderate	Rarely used $T_0^* = A^2T_b/4$
Noncoh. ASK	Same as above	$\simeq 2r_b$	$\frac{1}{2}\exp\left\{-\frac{A^2T_b}{16\eta}\right\}$	18.33	Minor	$T_0^* = A/2$ $P_{e0} \neq P_{e1}$
Coherent FSK	$s_1(t) = A\cos(\omega_c - \omega_d)t$ $s_2(t) = A\cos(\omega_c + \omega_d)t$ $2\omega_d = 1.5\pi r_b$	$>2r_b$	$Q\left(\sqrt{\frac{0.61A^2T_b}{\eta}}\right)$	10.6	Major	Seldom used; performance does not justify complexity $T_0^* = 0$
Noncoh. FSK	Same as above $2\omega_d = (k2\pi)r_b$	$>2r_b$	$\frac{1}{2}\exp\left\{-\frac{A^2T_b}{8\eta}\right\}$	15.33	Minor	Used for slow speed data transmission; poor utilization of power and bandwidth $T_0^* = 0.$
Coherent PSK	$s_1(t) = A\cos\omega_c t$ $s_2(t) = -A\cos\omega_c t$ $\omega_c = k2\pi r_b$	$\simeq 2r_b$	$Q\left(\sqrt{\frac{A^2T_b}{\eta}}\right)$	8.45	Major	Used for high speed data transmission. $T_0^* = 0$; best overall performance, but requires complex equipment
DPSK	Same as above with differential coding	$\simeq 2r_b$	$\frac{1}{2}\exp\left(-\frac{A^2T_b}{2\eta}\right)$	9.30	Moderate	Most commonly used in medium speed data transmission. $T_0^* = 0$; errors tend to occur in pairs

P_e—Prob. of error; A—carrier amplitude at receiver input; $\eta/2$—two-sided noise psd; T_b—bit duration; r_b—bit rate; $f_c = \omega_c/2\pi$ = carrier frequency; T_0^*—threshold setting; $S/N = A^2/2\eta r_b$; $P_{e0} = P$ (error|0 sent); $P_{e1} = P$ (error|1 sent).

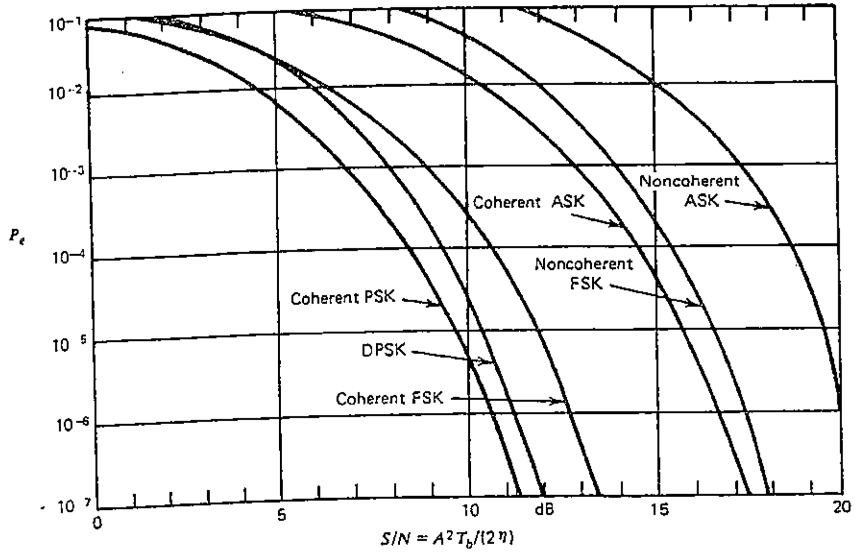


Figure 8.14 Probability of error for binary digital modulation schemes. (Note that the average signal power for ASK schemes is $A^2/4$, whereas it is $A^2/2$ for other schemes).

Plots in Figure 8.14 reveal that a coherent PSK signaling scheme requires the least amount of power followed by DPSK, coherent FSK, coherent ASK, noncoherent FSK, and noncoherent ASK signaling schemes. If the comparison is done in terms of average power requirements, then the ASK schemes require about the same amount of power as the FSK schemes. Since the cost of transmitting and receiving equipment depends more upon the peak power requirements than average power requirements, the comparison is usually made on the basis of peak power requirements. Thus, if the peak power requirement is of primary concern, then ASK schemes are not used.

It must be pointed out here that three of the most widely used digital modulation schemes are PSK, DPSK and noncoherent FSK. The power requirement of DPSK is approximately 1 dB more than coherent PSK, and the noncoherent FSK requires about 7 dB more power than coherent PSK. The reader may at this point ask the significance of say a 1 to 2 dB increase in power requirements. Industry sources claim that, in a large communication network, every 1 dB saving in power will result in savings of many millions of dollars annually.

8.6.3 Immunity to Channel Impairments

In selecting a signaling scheme, one should consider if the scheme is to some degree immune to channel impairments such as amplitude nonlinearities and fading (slow random variations in channel characteristics.) The FSK and PSK schemes are constant amplitude signals, and the threshold setting in the receiver does not depend on the received signal level. In the ASK scheme the receiver threshold setting depends on the received signal level and has to be changed as the received signal level changes. Thus ASK schemes are more sensitive to variations in received signal level due to changes in channel characteristics.

If the communication channel has fading, then noncoherent schemes have to be used because of the near impossibility of establishing a coherent reference at the receiver under fading channel conditions. However, if the transmitter has serious power limitations (as in the case of remote data transmission from space vehicles with limited energy storage and power generation capabilities), then a coherent scheme may have to be considered since coherent schemes use less power than noncoherent schemes for a given data rate and probability of error.

8.6.4 Equipment Complexity

There is very little difference in the complexity of transmitting equipment for the PSK, FSK, and ASK signals. At the receiver, the complexity depends on whether a coherent or noncoherent demodulation method is used. Hardware implementations of coherent demodulation schemes are more complex. Among the noncoherent schemes, DPSK is more complex than noncoherent FSK, which is more complex than noncoherent ASK. Complexity of equipment will increase the cost.

Summary. It must be obvious to the reader by now that there are a large number of factors that must be taken into account in the selection of a particular type of signaling scheme for a specific application. However, the following broad guidelines could be used to simplify the selection procedure:

1. If bandwidth is a premium quantity, then the most desirable signaling scheme is VSB with baseband signal shaping, and the least desirable scheme is FSK.
2. If power requirements are most important, then coherent PSK or DPSK is most desirable while ASK schemes are least desirable.

3. If equipment complexity is a limiting factor, then noncoherent demodulation schemes are preferrable to coherent schemes.

8.7 *M*-ARY SIGNALING SCHEMES

In Section 5.4 of Chapter 5, we saw that M-ary signaling schemes can be used for reducing the bandwidth requirements of baseband PAM data transmission systems. M-ary signaling schemes can be used in conjunction with digital modulation techniques also. Here, one of M $(M > 2)$ signals $s_1(t), s_2(t), \ldots, s_M(t)$ is transmitted during each signaling interval of duration T_s. These signals are generated by changing the amplitude, phase, or frequency of a carrier in M discrete steps. Thus we can have M-ary ASK, M-ary PSK, and M-ary FSK digital modulation schemes. M-ary digital modulation schemes are preferred over binary digital modulation schemes for transmitting digital information over bandpass channels when one wishes to conserve bandwidth (at the expense of increasing power requirements), or to conserve power (at the expense of increasing bandwidth requirements).

In practice, we seldom find a channel that has the exact bandwidth required for transmitting the output of a source using binary signaling schemes. When the bandwidth of the channel is less, M-ary digital modulation schemes are used to transmit the information over the bandpass channel. If the channel has a bandwidth much larger than the bandwidth required for transmitting the source output using binary modulation techniques, M-ary schemes may be used to utilize the additional bandwidth to provide increased immunity to channel noise. In this section, we will look at M-ary PSK schemes that are used for conserving bandwidth, and wideband M-ary FSK schemes that can be used for conserving power in digital modulation schemes.

In our discussion of M-ary schemes, we will assume that the input to the modulator is an independent sequence of equiprobable binary digits. We will further assume that the modulator takes blocks of λ binary digits and assigns one of M possible waveforms to each block $(M = 2^\lambda)$.

8.7.1 *M*-ary Coherent PSK

In M-ary PSK systems, the phase of the carrier is allowed to take on one of M possible values $\phi_k = k2\pi/M$ $(k = 0, 1, 2, \ldots, M-1)$. Thus the M possible signals that would be transmitted during each signaling interval of duration T_s are

$$s_k(t) = A\cos(\omega_c t + k2\pi/M), \quad k = 0, 1, \ldots, M-1, 0 \leq t \leq T_s \tag{8.51}$$

We will assume that ω_c, the carrier frequency, is a multiple of $r_s (r_s = 1/T_s)$. The digital M-ary PSK waveform can be represented in the form

$$Z(t) = A \sum_{k=-\infty}^{\infty} g(t - kT_s)\cos(\omega_c t + \phi_k) \tag{8.52}$$

where $g(t)$ is a rectangular unit amplitude pulse with a duration T_s. The sequence of phase angles $\{\phi_k\}$ carries the digital information. We can rewrite Equation (8.52) as

$$\begin{aligned} Z(t) = A\cos\omega_c t &\sum_{k=-\infty}^{\infty} (\cos\phi_k) g(t - kT_s) \\ - A\sin\omega_c t &\sum_{k=-\infty}^{\infty} (\sin\phi_k) g(t - kT_s) \end{aligned} \tag{8.53}$$

which shows that the waveform $Z(t)$ is the difference of two AM signals using $\cos\omega_c t$ and $\sin\omega_c t$ as carriers. The power spectral density of $Z(t)$ is a shifted version of the power spectral density of the M-ary rectangular waveforms $\Sigma\cos\phi_k g(t - kT_s)$ and $\Sigma\sin\phi_k g(t - kT_s)$. The psd of these waveforms has a $(\sin x/x)^2$ form with zero crossings at $\pm kr_s$ Hz. Thus the bandwidth requirement of an M-level PSK signal will be of the order of $2r_s$ to $3r_s$ Hz.

If the information to be transmitted is an independent binary sequence with a bit rate of r_b, then the bandwidth required for transmitting this sequence using binary PSK signaling scheme is of the order of $2r_b$. Now, if we take blocks of λ bits and use an M-ary PSK scheme with $M = 2^\lambda$ and $r_s = r_b/\lambda$, the bandwidth required will be of the order of $2r_s = 2r_b/\lambda$. Thus the M-ary PSK signaling scheme offers a reduction in bandwidth by a factor of λ over the binary PSK signaling scheme.

The M-ary PSK signal can be demodulated using a coherent demodulation scheme if a phase reference is available at the receiver. For purposes of illustration we will discuss the demodulation of four-phase PSK (also known as QPSK or quadrature PSK) in detail and then present the results for the general M-ary PSK.

In four-phase PSK, one of four possible waveforms is transmitted during each signaling interval T_s. These waveforms are:

$$\left.\begin{aligned} s_1(t) &= A\cos\omega_c t \\ s_2(t) &= -A\sin\omega_c t \\ s_3(t) &= -A\cos\omega_c t \\ s_4(t) &= A\sin\omega_c t \end{aligned}\right\} \text{for } 0 \leq t \leq T_s \tag{8.54}$$

These waveforms correspond to phase shifts of 0°, 90°, 180°, and 270° as shown in the phasor diagram in Figure 8.15. The receiver for the system is shown in Figure 8.16. The receiver requires two local reference waveforms

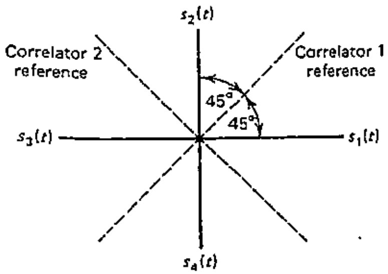


Figure 8.15 Phasor diagram for QPSK.

$A\cos(\omega_c t + 45°)$ and $A\cos(\omega_c t - 45°)$ that are derived from a coherent local carrier reference $A\cos\omega_c t$.

For purposes of analysis, let us consider the operation of the receiver during the signaling interval $(0, T_s)$. Let us denote the signal component at the output of the correlators by s_{01} and s_{02}, respectively, and the noise component by $n_0(t)$. If we assume that $s_1(t)$ was the transmitted signal during the signaling interval $(0, T_s)$, then we have

$$s_{01}(T_s) = \int_0^{T_s} (A\cos\omega_c t) A\cos\left(\omega_c t + \frac{\pi}{4}\right) dt$$

$$= \frac{A^2}{2} T_s \cos\frac{\pi}{4} = L_0$$

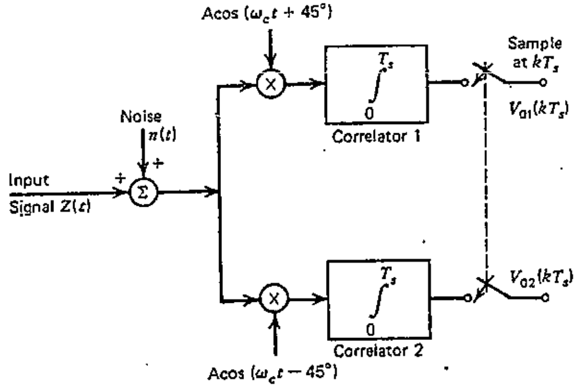


Figure 8.16 Receiver for QPSK scheme. Polarities of $V_{01}(kT_s)$ and $V_{02}(kT_s)$ determine the signal present at the receiver input during the kth signaling interval as shown in Table 8.4.

$$s_{02}(T_s) = \int_0^{T_s} (A\cos\omega_c t) A\cos\left(\omega_c t - \frac{\pi}{4}\right) dt$$

$$= \frac{A^2}{2} T_s \cos\frac{\pi}{4} = L_0$$

Table 8.4 shows s_{01} and s_{02} corresponding to each of the four possible signals $s_1(t)$, $s_2(t)$, $s_3(t)$, and $s_4(t)$.

Output signal levels shown in Table 8.4 indicate that the transmitted signal can be recognized from the polarities of the outputs of both correlators (i.e., the threshold levels are zero). In the presence of noise, there will be some probability that an error will be made by one or both correlators. An expression for the probability of incorrectly decoding the transmitted signal can be derived as follows.

The outputs of the correlators at time $t = T_s$ are

$$V_{01}(T_s) = s_{01}(T_s) + n_{01}(T_s)$$

$$V_{02}(T_s) = s_{02}(T_s) + n_{02}(T_s)$$

where $n_{01}(T_s)$ and $n_{02}(T_s)$ are zero mean Gaussian random variables defined by

$$n_{01}(T_s) = \int_0^{T_s} n(t) A\cos(\omega_c t + 45°)\, dt$$

$$n_{02}(T_s) = \int_0^{T_s} n(t) A\cos(\omega_c t - 45°)\, dt,$$

and $n(t)$ is a zero mean Gaussian random process with a power spectral density of $\eta/2$. With our assumption that $\omega_c = k2\pi r_s$ (k an integer > 0), we can show that $n_{01}(T_s)$ and $n_{02}(T_s)$ are independent Gaussian random variables with equal variance N_0 given by (see Problems 8.1, and 8.24)

$$N_0 = \frac{\eta}{4} A^2 T_s \tag{8.55}$$

Let us now calculate the probability of error assuming that $s_1(t)$ was the

Table 8.4. Output signal levels at sampling times.

	Input			
Output	$s_1(t)$	$s_2(t)$	$s_3(t)$	$s_4(t)$
$s_{01}(kT_s)$	L_0	$-L_0$	$-L_0$	L_0
$s_{02}(kT_s)$	L_0	L_0	$-L_0$	$-L_0$

transmitted signal. If we denote the probability that correlator 1 (Figure 8.16) makes an error by P_{ec1}, then

$$\begin{aligned} P_{ec1} &= P(n_{01}(T_s) < -L_0) \\ &= P(n_{01}(T_s) > L_0) \\ &= Q\left(\frac{L_0}{\sqrt{N_0}}\right) = Q\left(\sqrt{\frac{A^2T_s}{2\eta}}\right) \end{aligned} \tag{8.56}$$

By symmetry, the probability that the correlator 2 makes an error is

$$P_{ec2} = P_{ec1} = Q(\sqrt{A^2T_s/2\eta}) \tag{8.57}$$

The probability P_c that the transmitted signal is received correctly is

$$\begin{aligned} P_c &= (1 - P_{ec1})(1 - P_{ec2}) \\ &= 1 - 2P_{ec1} + P_{ec1}^2 \end{aligned}$$

We have made use of the fact that the noise outputs of the correlators are statistically independent at sampling times, and that $P_{ec1} = P_{ec2}$. Now, the probability of error P_e for the system is

$$\begin{aligned} P_e &= 1 - P_c \\ &= 2P_{ec1} - P_{ec1}^2 \\ &\approx 2P_{ec1} \end{aligned}$$

since P_{ec1} will normally be $\ll 1$. Thus for the QPSK system we have

$$P_e = 2Q(\sqrt{A^2T_s/2\eta}) \tag{8.58}$$

We can extend this result to the M-ary PSK signaling scheme when $M > 4$. In the general case of this scheme, the receiver consists of a phase discriminator—a device whose output is directly proportional to the phase of the incoming carrier plus noise measured over a signaling interval. The phase of the signal component at the receiver input is determined as θ_k if the phase discriminator output $\theta(t)$ at $t = kT_s$ is within $\pm\pi/M$ of θ_k. Thus the receiver makes an error when the magnitude of the noise-induced phase perturbation exceeds π/M (see Figure 8.17). A detailed derivation of an expression for the probability of error in an M-ary PSK scheme using ideal phase detection is given in References 1 and 8.

We will simply note here that the probability of error in an optimum M-ary PSK signaling scheme can be approximated by (see Reference 8, Chapter 14)

$$P_e \approx 2Q\left(\sqrt{\frac{A^2T_s}{\eta}\sin^2\frac{\pi}{M}}\right), \quad M \geq 4 \tag{8.59}$$

when the signal-to-noise power ratio at the receiver input is large. We are now ready to compare the power-bandwidth trade off when we use an M-ary PSK to transmit the output of a source emitting an independent sequence of equiprobable binary digits at a rate of r_b. We have already seen that if $M = 2^\lambda$ (λ an integer), the M-ary PSK scheme reduces the bandwidth by a factor of λ over the binary PSK scheme. It is left as an exercise for the reader to show that the ratio of the average power requirements of an M-ary PSK scheme $(S_{av})_M$ and the average power requirement of a binary PSK scheme $(S_{av})_b$ are given by

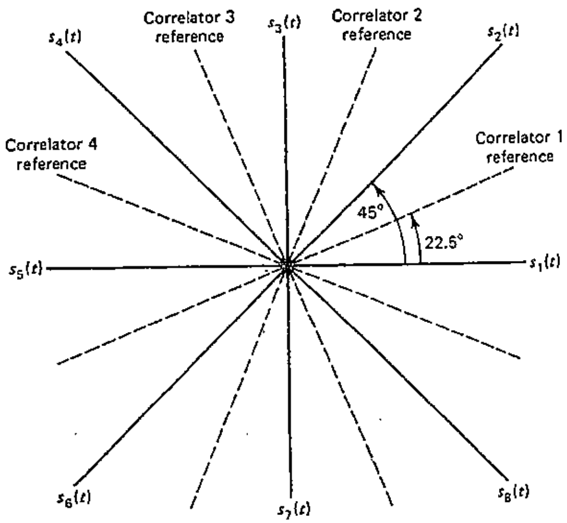


Figure 8.17 Phasor diagram for M-ary PSK; $M = 8$.

$$\frac{(S_{av})_M}{(S_{av})_b} = \left(\frac{z_1^2}{z_2^2}\right)\frac{1}{\lambda\sin^2(\pi/M)} \tag{8.60a}$$

where z_1 and z_2 satisfy $Q(z_1) = P_e/2$ and $Q(z_2) = P_e$, respectively. If P_e is very small, then z_1 will be approximately equal to z_2 and we can rewrite Equation (8.60a) as

$$\frac{(S_{av})_M}{(S_{av})_b} \approx \frac{1}{\lambda\sin^2(\pi/M)} \tag{8.60b}$$

Typical values of power bandwidth requirements for binary and M-ary schemes are shown in Table 8.5, assuming that the probability of error is equal to 10^{-4} and that the systems are operating in identical noise environments.

Table 8.5. Comparison of power-bandwidth requirements for M-ary PSK scheme. $P_e = 10^{-4}$.

Value of M	$\frac{(\text{Bandwidth})_M}{(\text{Bandwidth})_b}$	$\frac{(S_{av})_M}{(S_{av})_b}$
4	0.5	0.34 dB
8	0.333	3.91 dB
16	0.25	8.52 dB
32	0.2	13.52 dB

Values shown in the table indicate that the QPSK scheme offers the best trade off between power and bandwidth requirements. For this reason QPSK is very widely used in practice. For $M > 8$, power requirements become excessive and hence PSK schemes with $M > 8$ are very seldom used in practice. It must be pointed out here that M-ary PSK schemes require considerably more complex equipment than binary PSK schemes for signal generation and demodulation.

The results shown in Table 8.5 were derived under the assumption that the binary PSK and the M-ary PSK schemes operate with the *same symbol error probability* P_e. If the comparison is to be done with the same *bit error probability* P_{eb} for all schemes, then P_e should be modified according to Equation (5.57*a*) or (5.57*b*).

8.7.2 *M*-ary Differential PSK

M-ary PSK signals can be differentially encoded and demodulated using the phase comparison detection scheme discussed in Section 8.5. As an example of an M-ary differential PSK signaling scheme, let us consider the case where $M = 4$. The PSK signal with $M = 4$ given in Equation (8.53) can be thought of as two binary PSK signals using $\sin \omega_c t$ and $\cos \omega_c t$ as carriers. The four-phase PSK signal can be differentially encoded by encoding its two constituent binary PSK signals differentially, as explained in Table 8.2. The receiver for a four-phase differential PSK scheme consists of essentially two biphase comparison detectors, as shown in block diagram form in Figure 8.18.

Comparison of Figures 8.18 and 8.19 reveals that the receiver for the differential four-phase PSK signaling scheme uses a delayed version of the received signal as its phase reference. This principle was discussed earlier when we were discussing the binary DPSK signaling scheme in Section 8.4.2.

The performance of the four-phase differential PSK can be analyzed using a procedure that combines the results of Sections 8.4.2 and 8.5. The derivation of an expression for the probability of error for M-ary DPSK is rather involved and we will simply state the following expression (Reference 1) for the probability of error in an M-ary differential PSK scheme:

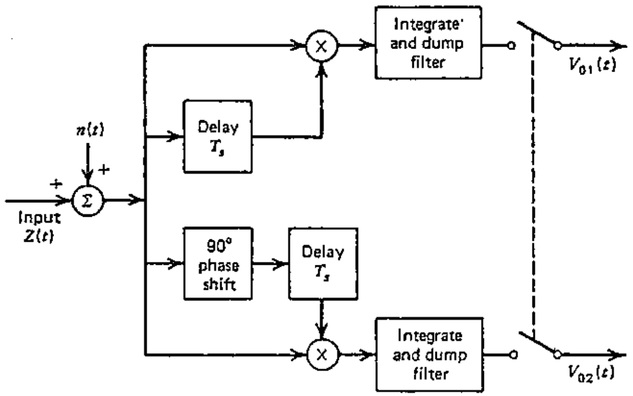


Figure 8.18 Receiver for four-phase differential PSK.

$$P_e = 2Q\left(\sqrt{\frac{A^2T_s}{\eta} 2\sin^2\left(\frac{\pi}{2M}\right)}\right) \tag{8.61}$$

Comparison of P_e for the M-ary DPSK scheme with P_e for the M-ary PSK scheme shows that differential detection increases power requirements by a factor of

$$\frac{\sin^2(\pi/M)}{2\sin^2(\pi/2M)} \approx 2 \text{ for large values of } M$$

With $M = 4$, the increase in power requirement is about 2 dB. This slight increase is more than offset by the simplicity of the equipment needed to handle the four-phase DPSK signal.

Figure 8.19 shows block diagrams of one of the earliest and most successful modems (modulator and demodulator) that uses a four-phase DPSK signaling scheme for transmitting data over (equalized) voice grade telephone lines. The binary data to be transmitted is grouped into blocks of two bits called *dibits*, and the resulting four possible combinations 00, 01, 10, and 11 differentially phase modulate the carrier. The data rate is fixed at 2400 bits/sec and the carrier frequency is 1800 Hz. The differential PSK waveform has the form

$$Z(t) = A\sum_k g(t - kT_s)\cos(\omega_c t + \phi_k)$$

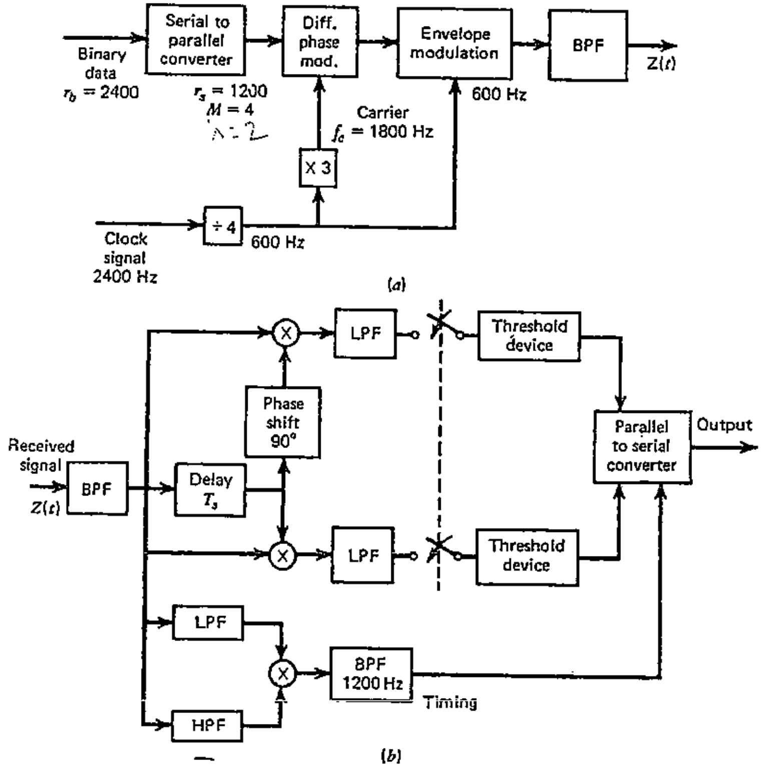


Figure 8.19 (*a*) Transmitter for differential PSK. (*b*) Receiver for differential PSK.

where

$$g(t) = \begin{cases} \frac{1}{2}(1 + \cos \pi r_s t) & \text{for } -T_s \leq t \leq T_s \\ 0 & \text{elsewhere, } T_s = 0.8333 \text{ msec} \end{cases}$$

The pulse shape $g(t)$ described above has negligible power content for $|f| > r_s$, thus the bandwidth of the transmitted signal is of the order of $2r_s$ or 2400 Hz. The non-rectangular shape for $g(t)$ conveys useful timing information to the receiver without introducing excessive ISI. Because of the differential phase-shift scheme (Table 8.6), the modulated waveform undergoes a phase change every T_s seconds. This feature along with the shape of $g(t)$ produces discrete frequency components at $f_c + 600$ and $f_c - 600$ that are used at the receiver to generate the 1200 Hz timing signal.

Table 8.6. Differential coding and decoding of quadrature PSK signals.

Dibit	$\phi_k - \phi_{k-1}$	$\sin(\phi_k - \phi_{k-1})$	$\cos(\phi_k - \phi_{k-1})$
00	+45°	+	+
01	+135°	+	−
10	−135°	−	−
11	−45°	−	+

The receiver produces two outputs—the first output is proportional to $\sin(\phi_k - \phi_{k-1})$ and the second is proportional to $\cos(\phi_k - \phi_{k-1})$. The input bits can be decoded uniquely from the sign of these two outputs as shown in Table 8.6. Finally, the parallel to serial converter interleaves the dibits to yield a serial binary output. Tests have shown that this modem has an error probability less than 10^{-5} when the signal-to-noise power ratio at the output of the channel is about 15 dB.

8.7.3 *M*-ary Wideband FSK Scheme

In this section we look at the possibility of using an M-ary FSK scheme to conserve power at the expense of (increased) bandwidth. Let us consider an FSK scheme where the M transmitted signals $s_i(t)$ $(i = 1, 2, \ldots, M)$ have the following properties:

$$s_i(t) = \begin{cases} A \cos \omega_i t, & 0 \leq t \leq T_s \\ 0 & \text{elsewhere} \end{cases} \tag{8.62}$$

and

$$\int_0^{T_s} s_i(t)s_j(t) = \begin{cases} A^2T_s/2 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{8.63}$$

The signals are of duration T_s, have equal energy, and are orthogonal to each other over the interval $(0, T_s)$. The minimum bandwidth of this signal set is approximately equal to $Mr_s/2$. In order to achieve this minimum bandwidth, the closest frequency separation $\omega_d = |\omega_n - \omega_m|$, $m \neq n$ must satisfy $\omega_d \geq \pi r_s$. One such choice of frequencies is $\omega_1 = k\pi r_s$, k an integer, and $\omega_m = \omega_1 + (m-1)\pi r_s$ $(m = 2, 3, \ldots, M)$.

The optimum receiver for this orthogonal signal set consists of a bank of M integrate and dump (or matched) filters as shown in Figure 8.20. The optimum receiver samples the filter output at $t = kT_s$ and decides that $s_i(t)$ was present at its input during the kth signaling interval if

$$\max_j [Y_j(kT_s)] = Y_i(kT_s) \tag{8.64}$$

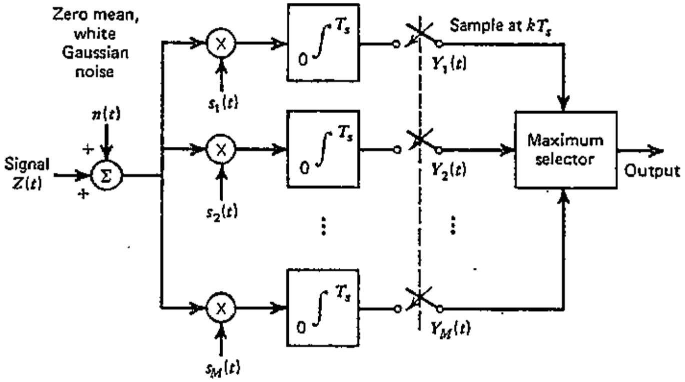


Figure 8.20 Structure of the receiver for an orthogonal (wideband FSK) signaling scheme.

To calculate the probability of error, let us consider the output of the filter at $t = T_s$ under the assumption that $s_1(t)$ was present at the receiver input during the interval $(0, T_s)$. The filter outputs are:

$$Y_j(T_s) = \int_0^{T_s} s_j(t)[n(t) + s_1(t)]\,dt, \quad j = 1, 2, \ldots, M \tag{8.65}$$

$$= \int_0^{T_s} s_j(t)s_1(t)\,dt + \int_0^{T_s} s_j(t)n't)\,dt \tag{8.66}$$

$$= s_{0j}(T_s) + n_j(T_s)$$

where $s_{0j}(T_s)$ is the signal component of the j-th filter output and $n_j(T_s)$ is the noise component. The reader can verify that (Problems 8.1 and 8.24)

$$s_{0j}(T_s) = \begin{cases} A^2T_s/2 & \text{if } j = 1 \\ 0 & \text{if } j \geqslant 2. \end{cases} \tag{8.67}$$

and $n_j(T_s)$ $(j = 1, 2, \ldots, M)$ are independent Gaussian random variables with zero means and equal variances N_0 given by

$$N_0 = A^2T_s(\eta/4) \tag{8.68}$$

The receiver correctly decodes that $s_1(t)$ was present during the signaling interval $(0, T_s)$ if $Y_1(T_s)$ is larger than $Y_j(T_s)$ $(j = 2, 3, \ldots, M)$. We calculate

this probability of correct decoding as

$$P_{c1} = P\{Y_2 < Y_1, Y_3 < Y_1, \ldots, Y_M < Y_1 | s_1 \text{ sent}\}$$

$$= \int_{-\infty}^{\infty} P\{Y_2 < y_1, \ldots, Y_M < y_1 | {}^{s_1 \text{ sent and}}_{Y_1 = y_1}\} f_{Y_1|s_1}(y_1)\,dy_1 \tag{8.69}$$

In the preceding step we made use of the identity

$$P(X < Y) = \int_{-\infty}^{\infty} P(X < y | Y = y)\, f_Y(y)\,dy$$

where X and Y are continuous, independent random variables.

We can simplify Equation (8.69) by making use of the fact that when s_1 is present during the signaling interval, $Y_2, Y_3, \ldots, Y_M$ are independent, zero mean Gaussian random variables with variance N_0, and hence the joint pdf of $Y_2, Y_3, \ldots, Y_M$ is given by

$$f_{Y_2, \ldots, Y_M | s_1; Y_1 = y_1}(y_2, \ldots, y_M) = \prod_{i=2}^{M} f_{Y_i}(y_i) \tag{8.70}$$

where

$$f_{Y_i}(y_i) = \frac{1}{\sqrt{2\pi N_0}} \exp\left(-\frac{y_i^2}{2N_0}\right), \quad -\infty < y_i < \infty$$

Substituting Equation (8.70) into (8.69), we have

$$P_{c1} = \int_{-\infty}^{\infty} \left\{ \underbrace{\int_{-\infty}^{y_1} \cdots \int_{-\infty}^{y_1}}_{M-1 \text{ integrals}} \prod_{i=2}^{M} f_{Y_i}(y_i)\,dy_i \right\} f_{Y_1|s_1}(y_1)\,dy_1$$

$$= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{y_1} f_Y(y)\,dy \right]^{M-1} f_{Y_1|s_1}(y_1)\,dy_1 \tag{8.71}$$

where

$$f_Y(y) = \frac{1}{\sqrt{2\pi N_0}} \exp\left(-\frac{y^2}{2N_0}\right), \quad -\infty < y < \infty$$

$$f_{Y_1|s_1}(y_1) = \frac{1}{\sqrt{2\pi N_0}} \exp\left(-\frac{(y_1 - s_{01})^2}{2N_0}\right), \quad -\infty < y_1 < \infty$$

and

$$N_0 = \left(\frac{A^2}{2}T_s\right)\frac{\eta}{2}$$
$$s_{01} = \frac{A^2}{2}T_s \tag{8.72}$$

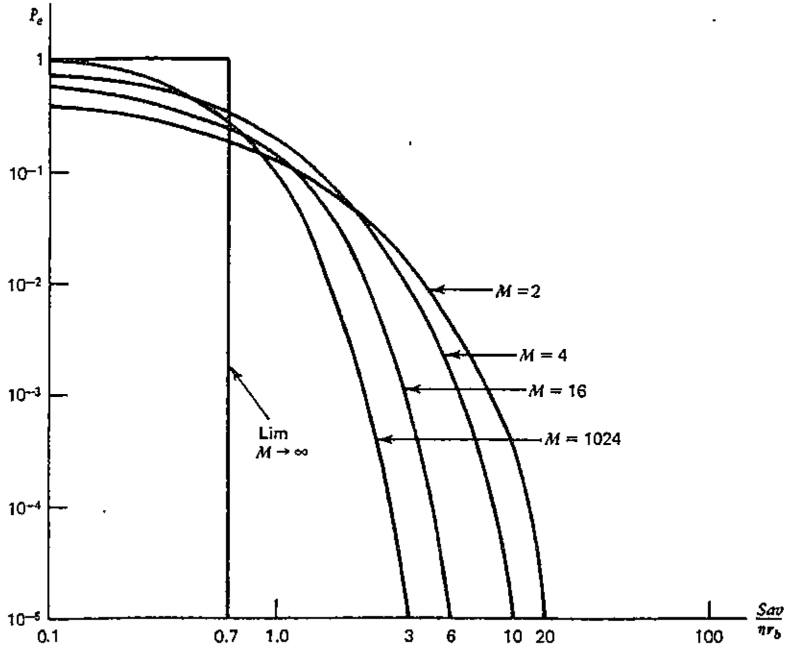


Figure 8.21 Probability of error for M-ary orthogonal signaling schemes.

Now, the probability that the receiver incorrectly decodes the incoming signal $s_1(t)$ is

$$P_{e1} = 1 - P_{c1}$$

and the probability that the receiver makes an error in decoding is

$$P_e = P_{e1}$$

because of symmetry.

The integral in Equation (8.71) cannot be expressed in a closed form for $M > 2$. Numerical integration techniques have been used to evaluate the integral and the results for several values of M are given in Figure 8.21.

The horizontal axis in Figure 8.21 is $S_{av}/\eta r_b$, where r_b is the data rate in bits/per second, S_{av} is the average signal power at the receiver input, and $\eta/2$ is the noise power spectral density at the receiver input. Because of our assumption that $M = 2^\lambda$, we have $r_b = r_s \log_2 M = \lambda r_s$ (λ a positive integer). The plots in Figure 8.21 reveal several interesting points. First, for fixed values of data rate r_b, noise psd $\eta/2$, and probability of error P_e, we see that

increasing values of M lead to smaller power requirements. Of course the price paid is the increase in bandwidth since the minimum bandwidth of M-ary orthogonal FSK signal set is $M/2T_s$ and it increases as the value of M increases. Also, large values of M lead to more complex transmitting and receiving equipment.

Figure 8.21 also reveals that in the limiting case as $M \to \infty$ the probability of error P_e satisfies

$$P_e = \begin{cases} 1 & \text{if } S_{av}/\eta r_b < 0.7 \\ 0 & \text{if } S_{av}/\eta r_b > 0.7 \end{cases}$$

The above relationship indicates that the maximum errorless rate r_b at which data can be transmitted using an M-ary orthogonal FSK signaling scheme is

$$r_b = \frac{S_{av}}{0.7\eta} \simeq \frac{S_{av}}{\eta} \log_2 e \tag{8.73}$$

The bandwidth of the signal set →∞ as $M \to \infty$.

It is interesting to note that the capacity C of a Gaussian channel of infinite bandwidth is $(S_{av}/\eta) \log_2 e$ (see Section 4.6). Equation (8.73) states that if the bit rate r_b is less than channel capacity, the probability of error can be made arbitrarily small. Thus we have indeed constructed a signaling scheme capable of signaling at a rate up to channel capacity with an arbitrarily small probability of error.

Example 8.6. Binary data is to be transmitted over a microwave channel at a rate of $(3)(10^6)$ bits/sec. Assuming the channel noise to be white Gaussian with a psd $\eta/2 = 10^{-14}$ watt/Hz, find the power and bandwidth requirements of four-phase PSK and 16-tone FSK signaling schemes to maintain an error probability of 10^{-4}.

Solution

(a) For the QPSK scheme we have

$$(P_e)_{\text{QPSK}} = 2Q(\sqrt{A^2T_s/2\eta})$$

where $T_s = 2T_b = (0.6667)10^{-6}$ and $\eta/2 = 10^{-14}$ watt/Hz, $P_e = 10^{-4}$, and hence

$$Q\left(\sqrt{\frac{A^2T_s}{2\eta}}\right) = \frac{10^{-4}}{2}$$

which requires

$$A^2T_s/2\eta = (3.9)^2$$

or

$$A^2/2 = S_{av} = -33.41 \text{ dBm}$$

The QPSK scheme requires a bandwidth of $2r_s = 3$ MHz.

(b) For the 16-tone FSK ($M = 16$), we obtain from Figure 8.21

$$S_{av}/\eta r_b = 5$$

or

$$S_{av} = (30)(10^{-8}) = -35.23 \text{ dBm}$$

for $P_e = 10^{-4}$. The bandwidth requirements of the 16-tone FSK scheme is $\geqslant M/2T_s$, where $M = 16$ and $T_s = (T_b) \log_2 16 = 4T_b = (1.333)(10^{-6})$. Hence the bandwidth required is $\geqslant 6$ MHz. Thus the multitone FSK has lower power requirements than QPSK, but requires more bandwidth and a more complex receiver structure.

8.8 SYNCHRONIZATION METHODS

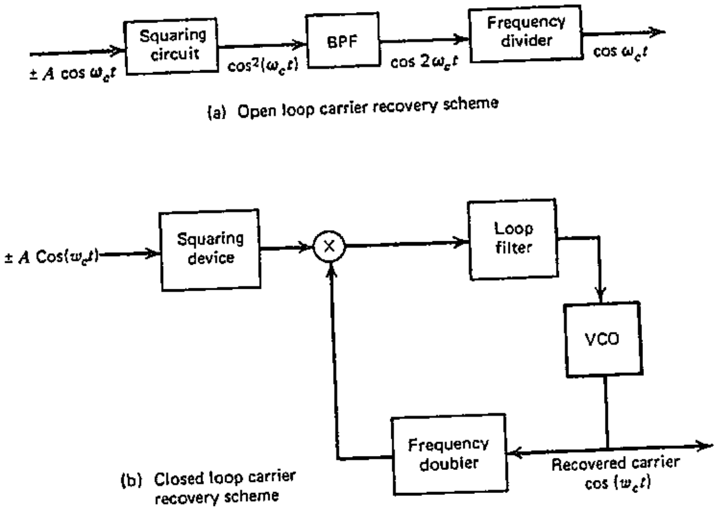
For optimum demodulation of ASK, FSK, and PSK waveforms, timing information is needed at the receiver. In particular, the integrate and dump operation in correlation receivers and the sampling operation in other types of receivers must be carefully controlled and sychronized with the incoming signal for optimum performance. Three general methods are used for synchronization in digital modulation schemes. These methods are:

1. Use of a primary or secondary time standard.
2. Utilization of a separate synchronization signal.
3. Extraction of clock information from the modulated waveform itself, referred to as self-synchronization.

In the first method, the transmitter and receiver are slaved to a precise master timing source. This method is often used in large data communication networks. In point-to-point data communication this method is very seldom used because of its cost.

Separate synchronization signals in the form of pilot tones are widely used in point-to-point data communication systems. In this method, a special synchronization signal or a sinusoidal signal of known frequency is transmitted along with the information carrying modulation waveform. The synchronization signal is sent along with the modulation waveform using one of the following methods:

1. by frequency division multiplexing, wherein the frequency of the pilot tone is chosen to coincide with a null in the psd of the signaling waveform;
2. by time division multiplexing where the modulated waveform (data stream) is interrupted for a short period of time during which the synchronizing signal is transmitted; or
3. by additional modulation such as the one shown in Figure 8.19.



(a) Open loop carrier recovery scheme

(b) Closed loop carrier recovery scheme

Figure 8.22 Extraction of local carrier for coherent demodulation of PSK signals.

In all of the above methods, the synchronization signal is isolated at the receiver and the zero crossings of the synchronization signal control the sampling operations at the receiver. All three methods discussed above add overhead (or additional requirements) to the system in terms of an increase in power and bandwidth requirements or a reduction in the data rate in addition to increasing the equipment complexity.

Self-synchronization methods extract a local carrier reference as well as timing information from the received waveforms. The block diagram of a system that derives a coherent local carrier from a PSK waveform is shown below in Figure 8.22a. Similar systems can be used to extract such a reference signal for other types of digital modulation schemes.

A feedback version of the squaring synchronizer is shown in Figure 8.22b. This version makes use of a PLL for extracting the correct phase and the frequency of the carrier waveform. The feedback version tracks the carrier phase more accurately, but its response is slower compared to the open-loop type synchronizing network.

If the carrier frequency is a multiple of the bit rate, then timing information can be derived from the local carrier waveform. Or, one could use self-synchronizing schemes similar to the ones described in Section 5.7.2. For these schemes to work properly, the incoming signal should have frequent symbol or bit changes. In some applications, the data stream might have to be scrambled at the transmitter to force frequent signal transitions (see Section 5.7.3).

8.9 SUMMARY

We developed procedures for analyzing and designing various signaling schemes for transmitting digital information over bandpass channels. Binary ASK, PSK, and FSK schemes were described in detail. Expressions for the probability of error for various schemes were derived in terms of average signal power at the receiver input, psd of the noise and signaling rate. The performances of various binary digital modulation schemes were compared. Finally, some of the commonly used M-ary signaling schemes were presented, and it was shown that a wideband M-ary orthogonal FSK scheme offers good trade-off between power and bandwidth.

It must be pointed out that combined modulation schemes, such as simultaneous amplitude and phase modulation, have also been used in data transmission applications. The treatment of combined modulation schemes is rather involved. The interested reader can find detailed treatment of these schemes in communication systems journals.

The signaling waveforms discussed in this chapter have spectral components that are nonzero for all values of frequencies—that is, the bandwidths of these signals approach infinity. In practical systems, filters are introduced to limit the transmission bandwidth. Such filtering introduces ISI and hence the performance of practical systems will be inferior to the performance of ideal systems discussed in this chapter. The analysis of systems that have ISI and additive noise is very complicated. Interested readers may refer to recent issues of the IEEE Transactions on Communications Technology which contain a large number of papers dealing with the combined effects of ISI and additive noise in digital communication systems.

REFERENCES

A very thorough treatment of digital modulation schemes, written at an advanced level, may be found in *Principles of Data Communications* by Lucky et al. (1968) and in *Modern Communication Principles* by Stein and Jones (1967). Sakrison's and Viterbi's books contain good theoretical treatment of signal detection theory and its application to digital modulation schemes. Many undergraduate texts provide easily readable treatments of several aspects of digital modulation schemes.

1. R. W. Lucky, J. Salz, and E. J. Weldon Jr. *Principles of Data Communications*. McGraw-Hill, New York (1968).
2. D. J. Sakrison. *Communication Theory: Transmission of Waveforms and Digital Information*. Wiley, New York (1968).
3. A. J. Viterbi. *Principles of Coherent Communication*. McGraw-Hill, New York, 1966.
4. Mischa Schwartz. *Information Transmission, Modulation and Noise*, 2nd ed. McGraw-Hill, New York (1970).
5. A. Bruce Carlson. *Communication Systems*. McGraw-Hill, New York (1975).
6. R. E: Ziemer and W. H. Tranter. *Principles of Communications*. Houghton Mifflin, Boston, Mass. (1976).
7. H. Taub and D. L. Schilling. *Principles of Communication Systems*. McGraw-Hill, New York (1971).
8. S. Stein and J. J. Jones. *Modern Communication Principles*. McGraw-Hill, New York (1967).

PROBLEMS

Section 8.2

8.1. $n(t)$ is a zero mean Gaussian white noise with a psd of $\eta/2$. $n_0(T_b)$ is related to $n(t)$ by

$$n_0(T_b) = \int_0^{T_b} n(t)s(t)\,dt$$

where $s(t) \equiv 0$ for t outside the interval $[0, T_b]$ and

$$\int_0^{T_b} s^2(t)\,dt = E_s$$

Show that $E\{n_0(T_b)\} = 0$ and $E\{[n_0(T_b)]^2\} = \eta E_s/2$.

8.2. A statistically independent sequence of equiprobable binary digits is transmitted over a channel having infinite bandwidth using the rectangular signaling waveform shown in Figure 8.23. The bit rate is r_b,

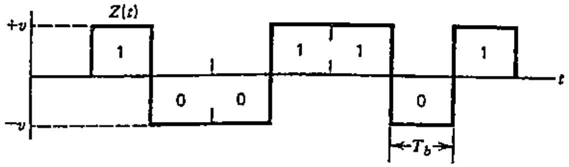


Figure 8.23 Signal waveform at the receiver input, Problem 8.2.

and the channel noise is white Gaussian with a psd of $\eta|2$.

(a) Derive the structure of an optimum receiver for this signaling scheme.

(b) Derive an expression for the probability of error.

8.3. In Problem 8.2, assume that the channel noise has a psd $G_n(f)$ given by

$$G_n(f) = G_0[1 + (f/f_1)^2]^{-1}$$

(a) Find the transfer function of the optimum receiver and calculate P_e.

(b) If an integrate and dump receiver is used instead of the optimum receiver, find P_e and compare with the P_e for the optimum receiver.

8.4. A received signal is ± 1 mv for T_b second intervals with equal probability. The signal is accompanied by white Gaussian noise with a psd of 10^{-10} watt/Hz. The receiver integrates the signal plus noise synchronously for T_b second duration and decodes the signal by comparing the integrator output with 0.

(a) Find the maximum signaling rate (minimum value of T_b) such that $P_e = 10^{-4}$.

(b) If actual signaling takes place at $\frac{1}{2}$ the rate found in (a), what is the signal amplitude required to maintain $P_e = 10^{-4}$?

8.5. Verify that the threshold value T_0^* shown in Figure 8.4 yields the minimum probability of error when $P(b_k = 0) = P(b_k = 1) = \frac{1}{2}$.

8.6. Assume that the ideal integrator in an integrate and dump receiver for Problem 8.2 is replaced by an RC lowpass filter with

$$H(f) = 1/(1 + jf/f_0)$$

where f_0 is the half-power frequency.

(a) Assuming the capacitor is initially discharged, find $s_{01}(T_b)$, $s_{02}(T_b)$, and $E\{n_0^2(t)\}$, where s_{01} and s_{02} are output signal values at $t = T_b$ and $n_0(t)$ is the noise at the filter output.

(b) Find the relationship between T_b and f_0 that will maximize $[s_{01}(T_b) - s_{02}(T_b)]^2/E\{n_0^2(t)\}$, that is, find the value of f_0 that will minimize P_e.

(c) Find the maximum value of $[s_{01} - s_{02}]^2/E\{n_0^2(t)\}$.

8.7. Referring to Equation (8.20), we have the signal-to-noise power ratio at the output of a matched filter receiver as

$$\gamma_{\max}^2 = \frac{2}{\eta}\int_0^T [s_1(t) - s_2(t)]^2\,dt$$

Now suppose that we want $s_1(t)$ and $s_2(t)$ to have the same signal energy. Show that the optimum choice of $s_2(t)$ is

$$s_2(t) = -s_1(t)$$

and that with $s_2(t) = -s_1(t)$

$$\gamma_{\max}^2 = (8/\eta)\int_0^T s_1^2(t)\,dt$$

8.8. An on-off binary system uses the following waveforms:

$$s_2(t) = \begin{cases} 2t/T_b, & 0 < t < T_b/2 \\ (2/T_b)(T_b - t), & T_b/2 \leq t < T_b \end{cases}$$

$$s_1(t) = 0$$

Assume that $T_b = 20\,\mu$sec and the noise psd is $\eta/2 = 10^{-7}$ watt/Hz. Find P_e for the optimum receiver assuming $P(0 \text{ sent}) = \frac{1}{4}$, $P(1 \text{ sent}) = \frac{3}{4}$.

Sections 8.3, 8.4, *and* 8.5

8.9. The input to a threshold device has the following conditional probabilities:

$$f_{R|0\text{ sent}}(r) = \frac{r}{N_0}\exp\left(-\frac{r^2}{2N_0}\right), \quad r > 0$$

$$f_{R|1\text{ sent}}(r) = \frac{1}{\sqrt{2\pi N_0}}\exp\left(-\frac{(r-A)^2}{2N_0}\right) \quad A \gg 0$$

$P(0 \text{ sent}) = \frac{1}{2}$ and $P(1 \text{ sent}) = \frac{1}{2}$. Find the optimum value of the threshold setting that will minimize the probability of error for $A = 1$ and $N_0 = 0.01$, 0.2, and 0.5. (*Hint*: Plot the pdf's and find the point where they intersect.) Compare the threshold values you get with the values obtained using the approximation given in Equation (8.37*b*).

8.10. In a binary PSK scheme using correlation receiver, the local carrier waveform is $A\cos(\omega_c t+\phi)$ instead of $A\cos\omega_c t$ due to poor carrier synchronization. Derive an expression for the probability of error and compute the increase in error probability when $\phi = 15°$ and $A^2T_b/\eta = 10$.

8.11. In a coherent binary PSK system, the peak carrier amplitude at the receiver A varies slowly due to fading. Assume that A has a pdf

$$f_A(a)=\frac{a}{A_0^2}\exp\left(-\frac{a^2}{2A_0^2}\right), \quad a \geqslant 0$$

(a) Find the mean and standard deviation of A.
(b) Find the *average* probability of error P_e. [Use the approximation for $Q(x)$ given in appendix D.]

8.12. In a coherent binary PSK system with $f_c = 5r_b$, the local carrier is in synchronism with the received signal, but the integrate and dump operation in the receiver is not fully synchronized. The sampling takes place at $t = 0.2T_b, 1.2T_b, 2.2T_b, \ldots$.
(a) How much intersymbol interference is generated by the offset in sampling times. (See Problem 5.8.)
(b) Calculate the probability of error and compare it with the probability of error that can be achieved with perfect timing.

8.13. In a coherent binary PSK system the symbol probabilities are $P(0 \text{ sent}) = p$ and $P(1 \text{ sent}) = 1-p$. The receiver is operating with a signal-to-noise ratio $(A^2T_b)/\eta = 4$.
(a) Find the optimum threshold setting for $p = 0.4, 0.5$, and 0.6 and find the probability of error P_e for $p = 0.4$, 0.5, and 0.6.
(b) Suppose that the receiver threshold setting was set at 0 for $p = 0.4$, 0.5, and 0.6. Find P_e and compare with P_e obtained in part (a).

8.14. An ideal analog bandpass channel has a usable bandwidth of 3000 Hz. The maximum average signal power allowed at the input to the channel is 0.001 mW. The channel noise (at the output) can be assumed to be zero mean white Gaussian with a psd of $\eta/2 = 10^{-10}$ watt/Hz.
(a) Find the capacity of the analog channel.
(b) Find the maximum rate at which binary data can be transmitted over this channel using binary PSK and FSK signaling schemes.
(c) Find P_e for coherent PSK and noncoherent FSK, assuming maximum signaling rate.
(d) Using the results of (b) and (c), find the capacity of the discrete channels corresponding to the coherent PSK and noncoherent FSK signaling schemes.

8.15. Consider a bandpass channel with the response shown in Figure 8.24.
(a) Binary data is transmitted over this channel at a rate of 300 bits/sec using a noncoherent FSK signaling scheme with tone frequencies of 1070 and 1270 Hz. Calculate P_e assuming $A^2/\eta = 8000$.
(b) How fast can a PSK signaling scheme operate over this channel? Find P_e for the PSK scheme assuming coherent demodulation.

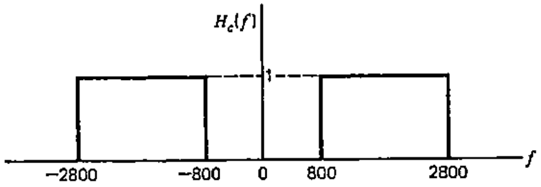


Figure 8.24 Response of a bandpass channel, Problem 8.15.

8.16. Compare the average power requirements of binary noncoherent ASK, coherent PSK, DPSK, and noncoherent FSK signaling schemes operating at a data rate of 1000 bits/sec over a bandpass channel having a bandwidth of 3000 Hz, $\eta/2 = 10^{-10}$ watt/Hz, and $P_e = 10^{-5}$.

8.17. Fill in the missing steps in the derivation of P_e for the DPSK signaling scheme.

8.18. A correlation receiver for a PSK system uses a carrier reference $A\sin\omega_c t$ for detecting

$$s_1(t) = A\cos(\omega_c t + \Delta\theta)$$

$$s_2(t) = A\sin(\omega_c t + \Delta\theta)$$

Assuming that $s_1(t)$ and $s_2(t)$ are equiprobable and the noise is white and Gaussian with a psd of $\eta/2$, find the probability of incorrect decoding.

8.19. The bit stream 11011100101 is to be transmitted using DPSK. Determine the encoded sequence and the transmitted phase sequence. Show that the phase comparison scheme described in Section 8.4.2 can be used for demodulating the signal.

8.20. A high frequency transmitter used in a binary communication system is peak power limited to 1 kW. The power loss in the channel is 60 dB and the noise power at the receiver input (ηr_b) is 10^{-4} watts. Assuming maximum signaling rate and equiprobable message bits, find P_e for noncoherent ASK and coherent PSK signaling schemes.

8.21. In some threshold devices a no-decision zone centered at the optimum threshold level is used such that if the input Y to the threshold device falls in this region, no decision is made, that is, the output is 0 if say $Y < T_1$ and 1 if $Y > T_2$ and no decision is made if $T_1 \leq Y \leq T_2$. Assuming that

$$f_{Y|1\text{ sent}}(y) = \frac{1}{2\sqrt{\pi}} \exp\left(-\frac{(y-1)^2}{4}\right), \quad -\infty < y < \infty$$

$$f_{Y|0\text{ sent}}(y) = \frac{1}{2\sqrt{\pi}} \exp\left(-\frac{(y+1)^2}{4}\right), \quad -\infty < y < \infty$$

$$P(1 \text{ sent}) = P(0 \text{ sent}) = 0.5$$

$$T_1 = -\epsilon, \quad T_2 = \epsilon, \quad \epsilon > 0$$

Sketch P_e and the probability of no decision versus ϵ. (Use $\epsilon = 0.1, 0.2, 0.3, 0.4,$ and 0.5.)

8.22. An ASK signaling scheme uses the noncoherent demodulation scheme shown in Figure 8.25. The center frequency of the filter is f_c and the bandwidth $B = 10r_b$ Hz. Assume that the bandwidth is such that the ASK signal passes through the filter with minimum distortion, and that the filter generates no ISI.

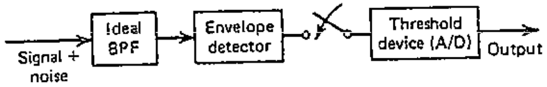


Figure 8.25 Noncoherent ASK receiver.

(a) Calculate the P_e for the receiver shown above assuming that $A^2/(\eta r_b) = 200$.

(b) Compare with P_e for a noncoherent ASK scheme if the filter is matched to the mark pulses.

8.23. Repeat Problem 8.22 for the noncoherent FSK scheme.

8.24. Let $n(t)$ be a stationary zero mean Gaussian white noise and let

$$n_{01}(T_b) = \int_0^{T_b} n(t) \cos(\omega_c t + \omega_d t)\, dt$$

$$n_{02}(T_b) = \int_0^{T_b} n(t) \cos(\omega_c t - \omega_d t)\, dt$$

Show that $n_{01}(T_b)$ and $n_{02}(T_b)$ are independent if $\omega_c = 2\pi k/T_b$ and $\omega_d = m\pi/2T_b$, where k and m are (arbitrary) positive integers $(k \gg m)$.

Section 8.6

8.25. An M-ary signaling scheme uses the following signals: $s_k(t) = A_k \cos(\omega_c t + \phi_k)$, $(0 \leq t \leq T_s)$, where

$$A_k = A \text{ or } 2A, \text{ and}$$

$$\phi_k = 45°, 90°, 135°, \text{ or } 270°,$$

(Observe that $M = 8$ and the signaling scheme is combined ASK/PSK.)

(a) Draw the block diagram of a coherent receiver for this system.

(b) Derive an approximate expression for the probability of error.

8.26. Consider the channel described in Problem 8.15.

(a) Compute the fastest rate at which data can be transmitted over this channel using four-phase PSK signaling schemes.

(b) Compute P_e for QPSK and differential QPSK.

8.27. A microwave channel has a usable bandwidth of 10 MHz. Data has to be transmitted over this channel at a rate of $(1.5)(10^6)$ bits/sec. The channel noise is zero mean Gaussian with a psd of $\eta/2 = 10^{-14}$ watt/Hz.

(a) Design a wideband FSK signaling scheme operating at $P_e = 10^{-5}$ for this problem, that is, find a suitable value of M and $A^2/2$.

(b) If a binary differential PSK signaling scheme is used for this problem, find its power requirement.

8.28. If the value of M obtained in Problem 8.27 is doubled, how will it affect the bandwidth and power requirements if P_e is to be maintained at 10^{-5}?

8.29. The design of a high-speed data communication system calls for a combined ASK/PSK signaling scheme with $M = 16$. Three alternate designs corresponding to three different sets of signaling waveforms are to be comparatively evaluated. The signaling waveforms in each set are shown in Figure 8.26 in a phasor diagram. The important parameters of the signal sets are the minimum distance between the phasors (parameter "a"

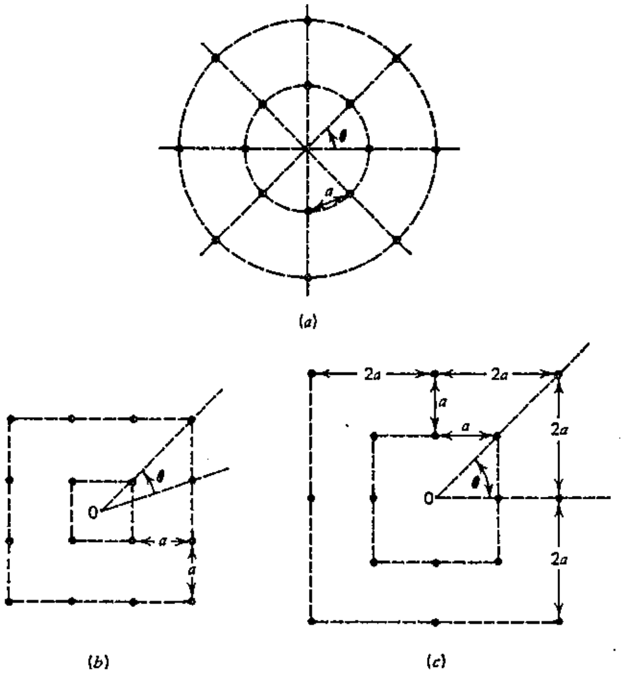


Figure 8.26 Constellations of signals, $M = 16$. Dots denote the tips of signal phasors.

in Figure 8.26 that is a measure of immunity against additive noise), the minimum phase difference (which is a measure of immunity against phase jitter/delay distortion), and the ratio of peak-to-average power (which is a measure of immunity against nonlinear distortion in the channel). Assuming that the average power is to be the same for all three signal sets, compare their robustness against the following channel impairments.

(a) Nonlinear distortion.
(b) Additive noise.
(c) Phase jitter.
(d) Combination of noise, phase jitter, and nonlinear distortion.

8.30. Derive the structure of a carrier recovery network (similar to the one shown in Figure 8.22a) for a QPSK signaling scheme.

9

ERROR CONTROL CODING

9.1 INTRODUCTION

In Chapters 5 and 8 we described signaling schemes for transmitting digital information over noisy channels. We saw that the probability of error for a particular signaling scheme is a function of the signal-to-noise ratio at the receiver input and the data rate. In practical systems the maximum signal power and the bandwidth of the channel are restricted to some fixed values due to governmental regulations on public channels or regulations imposed by private companies if the channel is leased. Furthermore, the noise power spectral density $\eta/2$ is also fixed for a particular operating environment. In addition, parameters of signaling schemes, such as the number and type of signals used, are chosen to minimize the complexity and cost of the equipment. With all of these constraints, it is often not possible to arrive at a signaling scheme which will yield an acceptable probability of error for a given application. Faced with this problem, the only practical alternative for reducing the probability of error is the use of error control coding, also known as channel coding.

In a nutshell, error control coding is the calculated use of *redundancy*. The functional blocks that accomplish error control coding are the channel encoder and the channel decoder. The channel encoder systematically adds digits to the transmitted message digits. These additional digits, while conveying no new information themselves, make it possible for the channel decoder to detect and correct errors in the information bearing digits. Error detection and/or correction lowers the overall probability of error.

Design a source encoding and a channel encoding scheme for this system that would yield an average symbol (letter) error rate of 50 symbols/sec. The source encoder is to use fixed length codewords with a block size ⩽ 3. Assume that one or more bit errors in the codeword will result in the incorrect decoding of all the symbols contained in the codeword (worst case).

10

DIGITAL TRANSMISSION OF ANALOG SIGNALS

10.1 INTRODUCTION

Communication systems are designed to handle the output of a variety of information sources. In the preceding chapters we considered analog communication systems using CW modulation schemes (AM, DSB, SSB, PM, and FM) for transmitting the output of analog information sources. We also discussed digital communication systems that used digital modulation schemes (discrete PAM, ASK, FSK, and PSK) for transmitting the output of discrete information sources. Simplified block diagrams of analog and digital communication systems are shown in Figure 10.1. In this chapter we will consider the use of digital communication systems such as the one shown in Figure 10.1*b* for transmitting the output of analog information sources.

Digital transmission of analog signals is possible by virtue of the sampling theorem which tells us that an analog signal can be reproduced from an appropriate set of its samples and hence we need transmit only the sample values as they occur rather than the analog signal itself. Samples of the analog signal can be transmitted using analog pulse modulation schemes wherein the amplitude, width, or position of a pulse waveform is varied in proportion to the values of the samples. The key distinction between analog pulse modulation and CW modulation is as follows: In CW modulation, some parameter of the modulated wave varies *continuously with the message.* In analog pulse modulation, some parameter of each pulse is modulated by a *particular sample value* of the message.

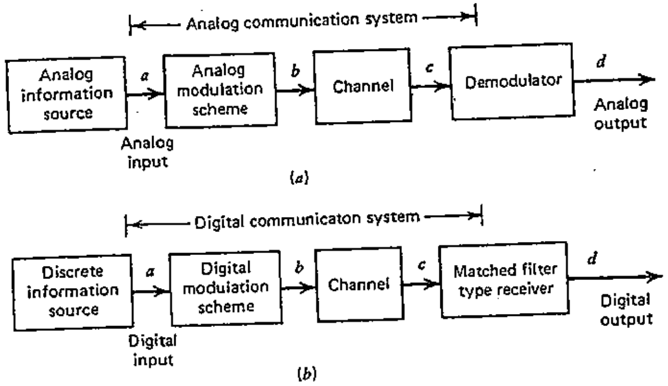

Analog communication system
Analog information source
a
Analog modulation scheme
b
Channel
c
Demodulator
d
Analog output
Analog input
(a)
Digital communicaton system
Discrete information source
a
Digital modulation scheme
b
Channel
c
Matched filter type receiver
d
Digital output
Digital input
(b)

Figure 10.1 Communication systems. (*a*) Analog communication system for transmitting the output of an analog information source. (*b*) Digital communication system for transmitting the output of a discrete information source.

Another method of transmitting the sampled values of an analog signal is to round off (quantize) the sampled values to one of Q predetermined values and then transmit the sampled and quantized signal using digital modulation schemes. The block diagram of a system that uses this scheme is shown in Figure 10.2. Here, the output $X(t)$ of the analog information source is converted to an M-ary symbol sequence $\{S_k\}$ through the processes of sampling, quantizing, and encoding. The M-ary sequence $\{S_k\}$ is transmitted using a digital communication system. The receiver output $\{\hat{S}_k\}$ will differ occasionally from the input $\{S_k\}$ due to transmission errors caused by channel noise and ISI in the digital communication system. An estimate $\hat{X}(t)$ of $X(t)$ is obtained from $\{\hat{S}_k\}$ through the process of decoding, and digital to analog (D/A) conversion. The reconstructed waveform $\hat{X}(t)$ will be a noisy version of the transmitted signal. The noise is due to sampling and quantizing of $X(t)$, and due to symbol errors that occur in the digital communication system. The overall performance of this system is measured by the signal-to-noise power ratio at the receiver output. A major portion of this chapter is devoted to the analysis of sampling, quantizing, and encoding procedures that are used to convert the analog output of an information source into a discrete symbol sequence suitable for transmission over a digital communication system.

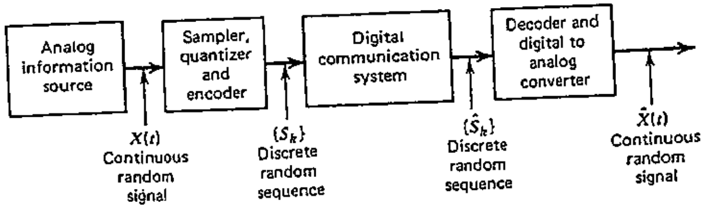


Figure 10.2 Digital transmission of analog signals.

At the beginning of this chapter, we will review the sampling theorem and discuss how analog signals are sampled and reconstructed in practical systems. We will then discuss methods of quantizing and encoding the sampled values for transmission over a digital communication system. Finally, we will derive expressions for the signal-to-noise power ratio at the receiver output and use these expressions for comparing the performance of digital and analog transmission schemes. We will also point out the advantages of using digital schemes for transmitting analog information.

In pulse communication systems, both analog and digital pulse modulation schemes are used. Analog pulse modulation, such as continuous pulse amplitude modulation and pulse position modulation, are similar to linear (AM) or exponential CW (PM or FM) modulation schemes. Digital or coded pulse modulation schemes such as pulse code modulation (PCM) and Delta modulation (DM) have no CW equivalent. We will treat only the digital pulse modulation schemes in this chapter. We begin our study with a review of sampling techniques.

10.2 SAMPLING THEORY AND PRACTICE

In many applications (such as in sample data control systems, digital computers, and in discrete pulse and CW modulation systems that we are currently dealing with) it is necessary and useful to represent an analog signal in terms of its sampled values taken at appropriately spaced intervals. In this section, we will first consider the representation of a low pass (bandlimited) deterministic signal $x(t)$ by its sampled values $x(kT_s)$ $(k = \ldots, -2, -1, 0, 1, 2, \ldots)$, where T_s is the time between samples. We will then extend the concept of sampling to include bandpass deterministic signals as well as to random signals. Finally, we will point out how the sampling and reconstruction of analog signals are carried out in practice.

10.2.1 Sampling Theory

The principle of sampling can be explained using the switching sampler shown in Figure 10.3. The switch periodically shifts between two contacts at a rate of

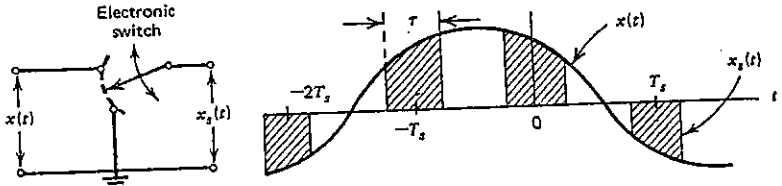


Figure 10.3 Switching sampler.

$f_s = 1/T_s$ Hz staying on the input contact for τ seconds and on the grounded contact for the remainder of each sampling period. The output $x_s(t)$ of the sampler consists of segments of $x(t)$, and $x_s(t)$ can be represented as

$$x_s(t) = x(t)s(t) \tag{10.1}$$

where $s(t)$ is the *sampling* or *switching function* shown in Figure 10.4.

Two questions that need to be answered with the sampling scheme shown in Figure 10.3 are: (1) Are the sampled segments sufficient to describe the original signal $x(t)$? (2) If so, how can we reconstruct $x(t)$ from $x_s(t)$? These questions can be answered by looking at the spectra (Fourier transforms) $X(f)$ and $X_s(f)$ of $x(t)$ and $x_s(t)$. Using the results derived in Chapter 2, we can express $s(t)$ as a Fourier series of the form

$$s(t) = C_0 + \sum_{n=1}^{\infty} 2C_n \cos n\omega_s t \tag{10.2}$$

where

$$C_0 = \tau/T_s, \quad C_n = f_s\tau \operatorname{sinc}[nf_s\tau], \quad \text{and} \quad \omega_s = 2\pi f_s$$

Combining Equations (10.2) and (10.1), we can write $x_s(t)$ as

$$x_s(t) = C_0x(t) + 2C_1x(t)\cos\omega_s t + 2C_2x(t)\cos 2\omega_s t + \cdots \tag{10.3}$$

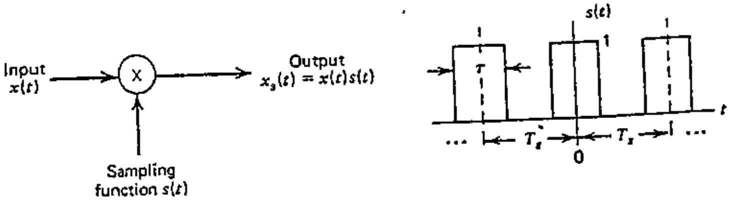


Figure 10.4 Sampling interpreted as multiplication. This type of sampling is often called *natural sampling.*

The Fourier transform of Equation (10.3) yields

$$X_s(f) = C_0X(f) + C_1[X(f-f_s) + X(f+f_s)] \\ + C_2[X(f-2f_s) + X(f+2f_s)] + \cdots \tag{10.4a}$$

$$= C_0X(f) + \sum_{\substack{n=-\infty \\ n\neq 0}}^{\infty} C_nX(f - nf_s) \tag{10.4b}$$

We can use Equation (10.4*a*) to find the spectrum of $x_s(t)$ given the spectrum of $x(t)$. Figure 10.5 shows the spectrum of the sampler output when the input $x(t)$ is bandlimited to f_x Hz.

It follows from Equation (10.4*a*) and from Figure (10.5*b*) that if $f_s > 2f_x$, then the sampling operation leaves the message spectrum intact, merely repeating it periodically in the frequency domain with a period of f_s. We also note that the first term in Equation (10.4*a*) (corresponding to the first term in Equation (10.3)) is the message term attenuated by the duty cycle C_0 of the sampling pulse. Since the sampling operation has not altered the message

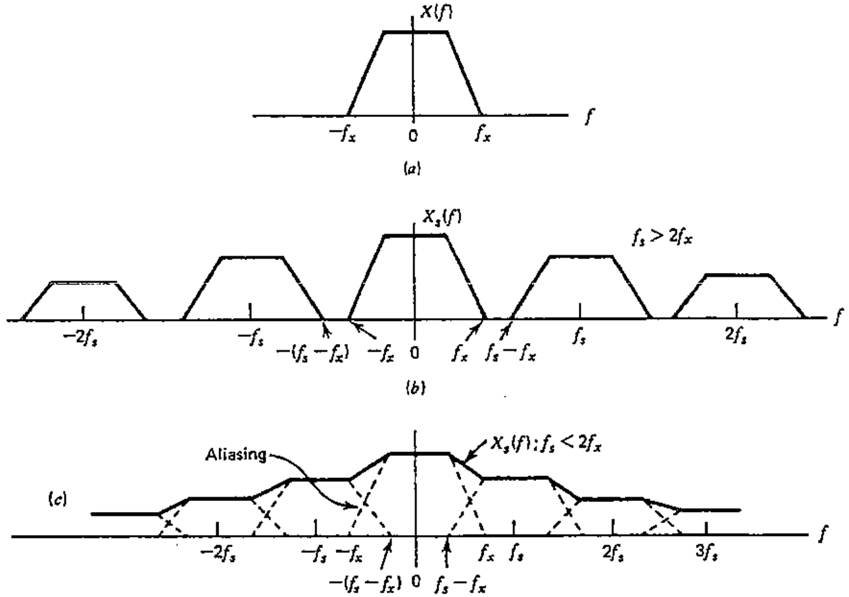


Figure 10.5 Sampling operation shown in frequency domain. (*a*) Message. (*b*) Sampled output, $f_s > 2f_x$. (*c*) Sampled output, $f_s < 2f_x$. (The width of the sampling pulse τ is assumed to be much smaller than T_s.)

spectrum, it should be possible to reconstruct $x(t)$ from the sampled waveform $x_s(t)$. While the procedure for reconstruction is not obvious from time domain relationships, it can be seen from Figure 10.5*b* that $X(f)$ can be separated from $X_s(f)$ by lowpass filtering. If we can filter $X(f)$ from $X_s(f)$, then we have recovered $x(t)$. Of course, such recovery is possible only when $x(t)$ is bandlimited and $f_s > 2f_x$. If the sampling rate $f_s < 2f_x$, then the sidebands of the signal overlap (Figure 10.5*c*) and $x(t)$ cannot be recovered without distortion from $X_s(f)$. This distortion is referred to as *aliasing*.

Thus the sampling frequency f_s must satisfy

$$f_s \geqslant 2f_x \quad \text{or} \quad T_s \leqslant 1/2f_x \tag{10.5}$$

The minimum sampling frequency $f_{s_{\min}} = 2f_x$ is called the *Nyquist rate.* When Equation (10.5) is satisfied, $x(t)$ can be recovered by passing $x_s(t)$ through an ideal lowpass filter with a bandwidth B, where B satisfies

$$f_x \leqslant B \leqslant f_s - f_x \tag{10.6}$$

At this point, we restate our reason for studying sampling theory: namely, we want to represent an analog signal by a sequence of sampled values. So far we have seen how an analog signal can be represented by a sequence of *segments*; now we proceed to show that indeed it is sufficient to have *instantaneous values* of $x(t)$ rather than segments of $x(t)$ for adequate representation of $x(t)$.

10.2.2 Ideal Sampling and Reconstruction of Lowpass Signals

Ideal sampling, by definition, is *instantaneous sampling*, and is accomplished by using a train of impulses $s_\delta(t)$ as the sampling function. Thus we have, for ideal sampling,

$$x_\delta(t) = x(t)s_\delta(t) \tag{10.7}$$

where

$$s_\delta(t) = \sum_{k=-\infty}^{\infty} \delta(t - kT_s) \tag{10.8}$$

Using the properties of the uniformly spaced impulse train, the reader can verify that

$$\begin{aligned} x_\delta(t) &= x(t) \sum_{k=-\infty}^{\infty} \delta(t - kT_s) \\ &= \sum_{k=-\infty}^{\infty} x(kT_s)\delta(t - kT_s) \end{aligned} \tag{10.9}$$

and

$$X_\delta(f) = X(f) * S_\delta(f) \tag{10.10}$$

where

$$S_\delta(f) = f_s \sum_{k=-\infty}^{\infty} \delta(f - nf_s) \tag{10.11}$$

or

$$X_\delta(f) = f_s \sum_{n=-\infty}^{\infty} X(f - nf_s) \tag{10.12}$$

Comparing Equation (10.12) and (10.4*b*), we see that the only difference is that the constants C_n in Equation (10.4*b*) are equal to f_s in Equation (10.12). Thus for perfect reconstruction of $x(t)$ from $x_\delta(t)$ we invoke the same conditions as we had for recovering $x(t)$ from $x_s(t)$, that is, $x(t)$ must be bandlimited to f_x and $f_s \geqslant 2f_x$. Then, we can reconstruct $x_\delta(t)$ from $x(t)$ by passing $x_\delta(t)$ through an ideal lowpass filter $H_R(f)$ with a bandwidth B satisfying $f_x \leqslant B \leqslant f_s - f_x$ as shown in Figure 10.6. Next, if we let the filter gain $K = 1/f_s$, then, from Equation (10.12) and Figure 10.6, we have

$$X(f) = X_\delta(f)H_R(f)$$

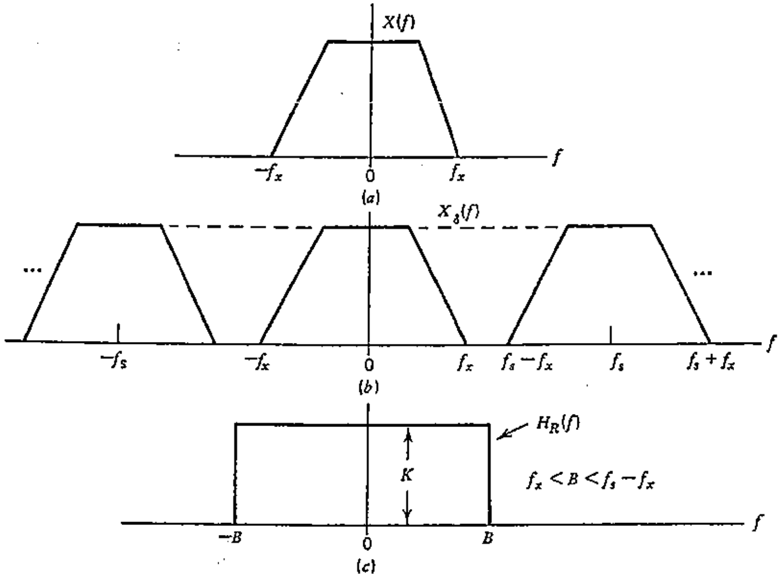


Figure 10.6 Spectra of ideally sampled signals. (*a*) Message. (*b*) Ideally sampled message. (*c*) Reconstruction filter. $X_\delta(f)H_R(f) = CX(f)$, where $C = Kf_s$.

In the time domain, we can represent the reconstruction process by

$$x(t) = F^{-1}\{X_\delta(f)H_R(f)\} \\ = [h_R(t) * x_\delta(t)] \tag{10.13}$$

where $h_R(t)$ is the impulse response of the reconstruction filter. We know that $h_R(t)$ is $2BT_s$ sinc$(2Bt)$ and hence we have

$$x(t) = [2BT_s \operatorname{sinc}(2Bt)] * \left[\sum_{k=-\infty}^{\infty} x(kT_s)\delta(t-kT_s)\right] \\ = 2BT_s \sum_{k=-\infty}^{\infty} x(kT_s) \operatorname{sinc} 2B(t-kT_s) \tag{10.14}$$

Equation (10.14) gives us the result we were seeking; namely, a bandlimited signal $x(t)$ can be represented by a sequence of sampled values $\{x(kT_s)\}$ if the sampling is done such that $f_s > 2f_x$. We state this result in the following theorem:

The Uniform Sampling Theorem for Lowpass Signals

If a signal $x(t)$ contains no frequency components for $|f| > f_x$, then it is completely described by instantaneous values $x(kT_s)$ uniformly spaced in time with period $T_s \leq 1/2f_x$. If the sampling rate f_s is equal to the Nyquist rate or greater ($f_s \geq 2f_x$), and if the sampled values are represented by weighted impulses, then the signal can be exactly reconstructed from its samples by an ideal lowpass filter of bandwidth B, where $f_x \leq B \leq f_s - f_x$ and $f_s = 1/T_s$.

10.2.3 Ideal Sampling and Reconstruction of Bandpass Signals

Signals with bandpass spectra can also be represented by their sampled values. Consider a signal $x(t)$ with the spectrum shown in Figure 10.7. The following sampling theorem gives the conditions for representing $x(t)$ by its sampled values.

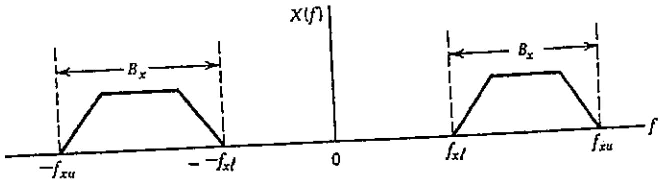


Figure 10.7 Spectrum of a bandpass signal.

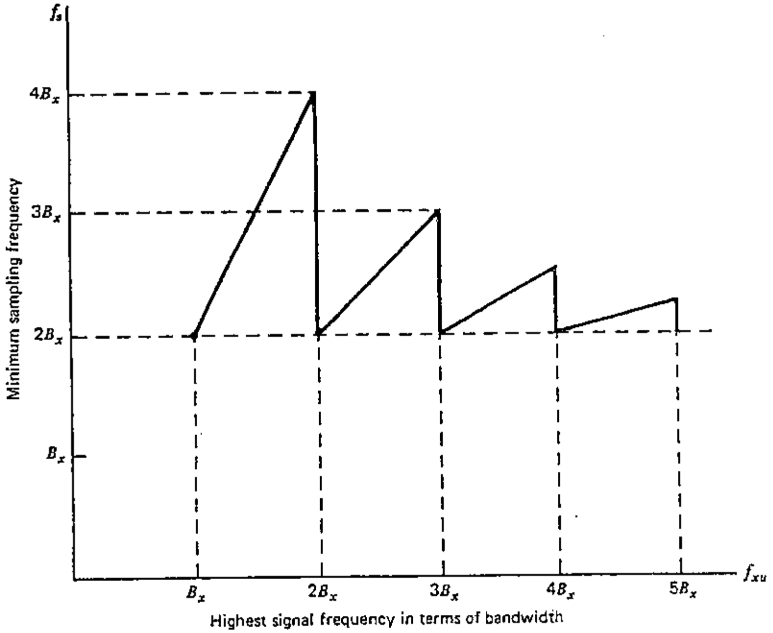


Figure 10.8 Minimum sampling frequency for a signal occupying a bandwidth B_x.

The Uniform Sampling Theorem for Bandpass Signals

If a bandpass signal $x(t)$ has a spectrum of bandwidth B_x and upper frequency limit f_{xu}, then $x(t)$ can be represented by instantaneous values $x(kT_s)$ if the sampling rate f_s is $2f_{xu}/m$, where m is the largest integer not exceeding f_{xu}/B_x. (Higher sampling rates are not always usable unless they exceed $2f_{xu}$.) If the sample values are represented by impulses, then $x(t)$ can be exactly reproduced from its samples by an ideal bandpass filter $H(f)$ with the response

$$H(f) = \begin{cases} 1 & f_{xl} < |f| < f_{xu} \\ 0 & \text{elsewhere} \end{cases}$$

The sampling rate for a bandpass signal depends on the ratio f_{xu}/B_x. If $f_{xu}/B_x \gg 1$, then the minimum sampling rate approaches $2B_x$. A sketch of f_{xu}/B_x versus f_s/B_x is shown in Figure 10.8. The reader can easily verify that $f_s > 2f_{xu}$ will result in exact reconstruction. Proof of exact reconstruction when

$f_s = 2f_{xu}/m$, where m is an integer satisfying $(f_{xu}/B_x) - 1 < m \leq f_{xu}/B_x$, is left as an exercise for the reader (see Problems 10.3 and 10.4).

10.2.4 Sampling Theorem for Random Signals

Having looked at sampling methods for deterministic signals, we now turn our attention to the sampling of random processes. The message waveform $X(t)$ in communication systems is often modeled as a bandlimited stationary random process at the baseband level. The power spectral density $G_X(f)$ of a bandlimited random process $X(t)$ is zero for $|f| > f_x$. Hence the autocorrelation function $R_{XX}(\tau)$ can be written as (see Equation (10.14))

$$R_{XX}(\tau) = 2BT_s \sum_{k=-\infty}^{\infty} R_{XX}(kT_s) \operatorname{sinc} 2B(\tau - kT_s) \tag{10.15}$$

where $1/T_s = f_s > 2f_x$ and $f_x < B < f_s - f_x$. It is convenient to state two different versions of Equation (10.15). With a an arbitrary constant, the transform of $R_{XX}(\tau - a)$ is equal to $G_X(f) \exp(-2\pi j f a)$. This function is also bandlimited, and hence Equation (10.15) can be applied to $R_{XX}(\tau - a)$ as

$$R_{XX}(\tau - a) = 2BT_s \sum_{n=-\infty}^{\infty} R_{XX}(nT_s - a) \operatorname{sinc} 2B(\tau - nT_s) \tag{10.16}$$

Changing $(\tau - a)$ to τ in Equation (10.16), we have

$$R_{XX}(\tau) = 2BT_s \sum_{n=-\infty}^{\infty} R_{XX}(nT_s - a) \operatorname{sinc} 2B(\tau + a - nT_s) \tag{10.17}$$

We will now state and prove the sampling theorem for bandlimited random processes using Equations (10.15) and (10.16).

The Uniform Sampling Theorem for Bandlimited Random Signals

If a random process $X(t)$ is bandlimited to f_x Hz, then $X(t)$ can be represented using the instantaneous values $X(kT_s)$ as

$$X(t) \overset{MS}{=} \hat{X}(t) = 2BT_s \sum_{-\infty}^{\infty} X(nT_s) \operatorname{sinc}[2B(t - nT_s)] \tag{10.18}$$

(where $\overset{MS}{=}$ stands for equality in the mean squared sense*) if the sampling rate f_s is equal to or greater than the Nyquist rate $2f_x$. If the sampled values are represented by weighted impulses, then $X(t)$ can be reconstructed from its samples by an ideal lowpass filter of bandwidth B, where $f_x \leq B \leq f_s - f_x$ and $f_s = 1/T_s$.

*$X(t) \overset{MS}{=} \hat{X}(t)$ if $E\{[X(t) - \hat{X}(t)]^2\} = 0$.

To prove Equation (10.18), we need to show that

$$E\{[X(t) - \hat{X}(t)]^2\} = 0 \tag{10.19}$$

where

$$\hat{X}(t) = 2BT_s \sum_{n=-\infty}^{\infty} X(nT_s) \operatorname{sinc}[2B(t - nT_s)]$$

Now,

$$\begin{aligned} E\{[X(t) - \hat{X}(t)]^2\} = {} & E\{[X(t) - \hat{X}(t)]X(t)\} \\ & - E\{[X(t) - \hat{X}(t)]\hat{X}(t)\} \end{aligned} \tag{10.20}$$

The first term on the right-hand side of the previous equation may be written as

$$\begin{aligned} & E\{[X(t) - \hat{X}(t)]X(t)\} \\ & \quad = R_{XX}(0) - 2BT_s \sum_{n=-\infty}^{\infty} R_{XX}(nT_s - t) \operatorname{sinc}[2B(t - nT_s)] \end{aligned}$$

From Equation (10.17) with $\tau = 0$ and $a = t$, we have

$$2BT_s \sum_{n=-\infty}^{\infty} R_{XX}(nT_s - t) \operatorname{sinc}[2B(t - nT_s)] = R_{XX}(0)$$

and hence

$$E\{[X(t) - \hat{X}(t)]X(t)\} = 0 \tag{10.21}$$

The second term in Equation (10.20) can be written as

$$\begin{aligned} & E\{[X(t) - \hat{X}(t)]\hat{X}(t)\} \\ & \quad = \sum_{m=-\infty}^{\infty} E\{[X(t) - \hat{X}(t)]X(mT_s)\} 2BT_s \operatorname{sinc}[2B(t - mT_s)] \end{aligned}$$

Now,

$$\begin{aligned} & E\{[X(t) - \hat{X}(t)]X(mT_s)\} \\ & \quad = R_{XX}(t - mT_s) - \sum_{n=-\infty}^{\infty} 2BT_s R_{XX}(nT_s - mT_s) \operatorname{sinc}[2B(t - nT_s)] \end{aligned}$$

and from Equation (10.16) with $\tau = t$ and $a = mT_s$, we have

$$R_{XX}(t - mT_s) = 2BT_s \sum_{n=-\infty}^{\infty} R_{XX}(nT_s - mT_s) \operatorname{sinc}[2B(t - nT_s)]$$

Hence,

$$E\{[X(t) - \hat{X}(t)]\hat{X}(t)\} = 0 \tag{10.22}$$

Substitution of Equations (10.21) and (10.22) in (10.20) completes the proof of Equation (10.19).

The proof of the second part of the theorem dealing with the reconstruction follows the steps outlined in Section 10.2.2. If the random process $X(t)$ is a bandpass process, then a theorem similar to the uniform sampling theorem for deterministic bandpass signals can be developed.

The sampling theorems for random signals tell us that the output of analog information sources can be adequately represented by the sampled values of the signals. Thus, rather than transmitting an analog signal, we need to transmit only the sampled values. At the receiver, the analog signal can be reconstructed from the received sequence of sampled values by appropriate filtering.

10.2.5 Practical Sampling

There are a number of differences between the ideal sampling and reconstruction techniques described in the preceding sections and the actual signal sampling as it occurs in practice. The major differences are:

1. The sampled wave in practical systems consists of finite amplitude and finite duration pulses rather than impulses.
2. Reconstruction filters in practical systems are not ideal filters.
3. The waveforms that are sampled are often timelimited signals and hence are not bandlimited.

Let us look at the effects of these differences on the quality of the reconstructed signals.

The sampled waveform produced by practical sampling devices, especially the sample and hold variety, has the form

$$\begin{aligned} x_s(t) &= \sum_{k=-\infty}^{\infty} x(kT_s)p(t-kT_s) \\ &= [p(t)] * \left[\sum_{k=-\infty}^{\infty} x(kT_s)\delta(t-kT_s)\right] \end{aligned}$$

where $p(t)$ is a flat topped pulse of duration τ. (This type of sampling is called *flat topped sampling*.) The spectrum $X_s(f)$ of $x_s(t)$ is given by

$$X_s(f) = P(f)X_\delta(f) = P(f)\left[f_s \sum_{n=-\infty}^{\infty} X(f-nf_s)\right] \tag{10.23}$$

where $P(f)$ is the Fourier transform of $p(t)$ and $X_\delta(f)$ is the Fourier transform of the ideal sampled wave. $P(f)$ is a sinc function and hence we can say from Equation (10.23) that the primary effect of flat topped sampling is an attenuation of high-frequency components. This effect, sometimes called an *aperture effect*, can be compensated by an equalizing filter with a transfer function $H_{eq}(f) = 1/P(f)$. However, if the pulsewidth is chosen to be small compared

to the time between samples (i.e., $\tau \ll 1/f_x$) then $P(f)$ is essentially constant over the message band and no equalization may be needed. Thus, effects of pulse shape are often unimportant and flat topped sampling is a good approximation to ideal impulse sampling.

The effect of nonideal reconstruction filters is shown in Figure 10.9. To recover the sampled signal shown in the figure, we need an ideal lowpass filter. However, such filters can only be approximated in practice. The output of the filter shown in Figure 10.9*a* will consist of $x(t)$ plus spurious frequency components at $|f| > f_x$ that lie outside the message band. While these components are considerably attenuated compared to $x(t)$, their presence may be annoying in some applications such as in audio systems. Good filter design will minimize this effect. Alternatively, for a given filter response, the high-frequency spurious components can be suppressed or eliminated by increasing the sampling frequency as shown in Figure 10.9*b*. Increasing the sampling frequency produces a *guard band* of width $(f_s - 2f_x)$ Hz.

In many practical applications, the waveform to be sampled might last for only a finite amount of time. Such message waveforms are not strictly bandlimited, and when such a message is sampled, there will be unavoidable overlapping of spectral components at frequencies $f > f_s/2$ (see Figure 10.10). The effect of this overlapping (also called aliasing) is far more serious than spurious high-frequency components passed by nonideal reconstruction filters, for the latter fall outside the message band. Aliasing effects can be minimized by bandlimiting the signal by filtering before sampling and sampling at a rate moderately higher than the nominal Nyquist rate.

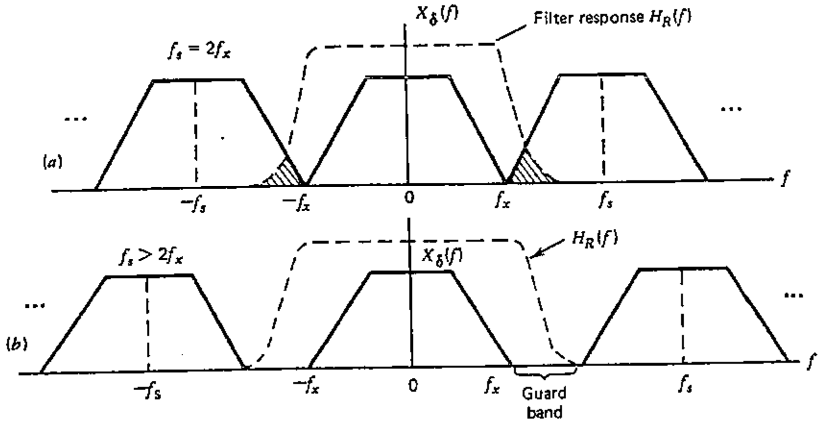


Figure 10.9 Reconstruction of sampled signals.

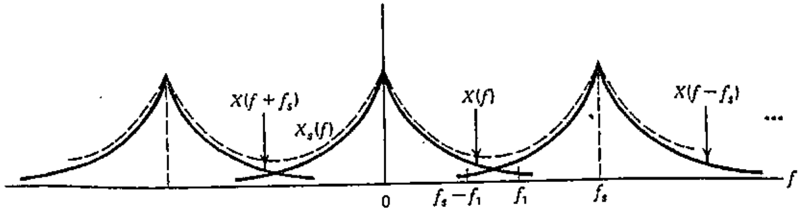


Figure 10.10 Sampling of nonbandlimited signals.

10.3 QUANTIZING OF ANALOG SIGNALS

In the preceding sections we established the fact that an analog message signal can be adequately represented by its sampled values. Message signals such as speech waveforms or video waveforms have a continuous amplitude range and hence the samples are also continuous in amplitude. When these continuous amplitude samples are transmitted over a noisy channel, the receiver cannot discern the exact sequence of transmitted values. The effect of the noise in the system can be minimized by representing the samples by a finite number of predetermined levels and transmitting the levels using a discrete signaling scheme such as discrete PAM. Now, if the separation between the levels is large compared to the noise perturbations, it will be a simple matter for the receiver to decide precisely which specific value was transmitted. Thus the effect of random noise can be virtually eliminated.

Representing the analog sampled values by a finite set of levels is called *quantizing*. While sampling converts a continuous time signal to a discrete time signal, quantizing converts a continuous amplitude sample to a discrete amplitude sample. Thus sampling and quantizing operations convert the output of an analog information source into a sequence of levels (or symbols), that is, the analog source is transformed to a discrete (digital) source. The sequence of levels can be transmitted using any one of the many digital signaling schemes discussed in the preceding chapters. An example of the quantizing operation is shown in Figure 10.11.

The input to the quantizer is a random process $X(t)$ that represents the output of an analog information source. The random waveform $X(t)$ is sampled at an appropriate rate and the sampled values $X(kT_s)$ are converted to one of Q allowable levels, $m_1, m_2, \ldots, m_Q$, according to some predetermined rule:

$$X_q(kT_s) = m_i \quad \text{if} \quad x_{i-1} \leqslant X(kT_s) < x_i \qquad (10.24)$$
$$x_0 = -\infty,\ x_Q = +\infty$$

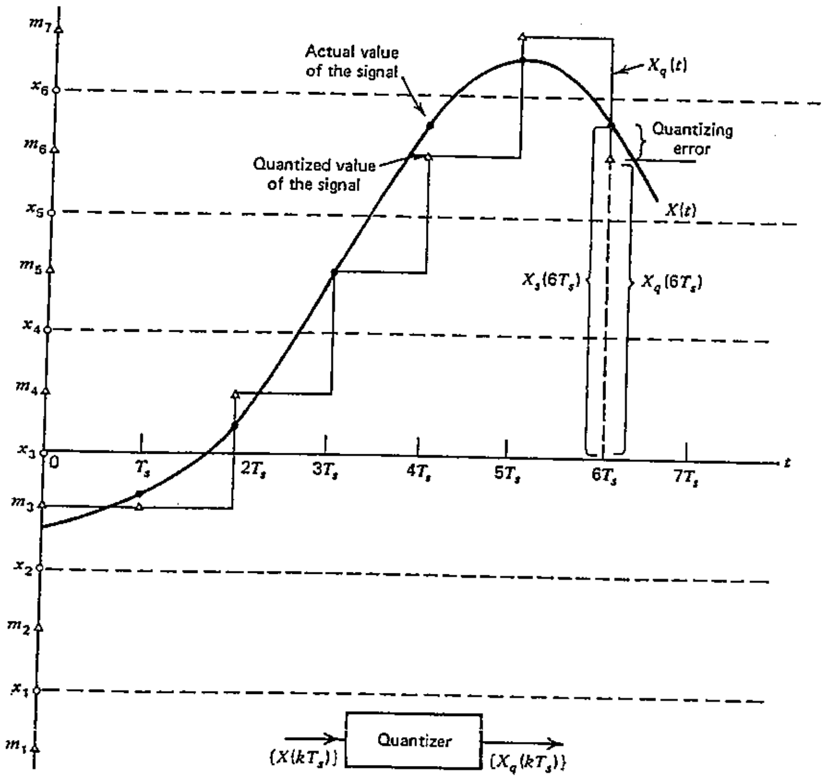


Figure 10.11 Quantizing operation; $m_1, m_2, \ldots, m_7$ are the seven output levels of the quantizer.

The output of the quantizer is a sequence of levels, shown in Figure 10.11 as a waveform $X_q(t)$, where

$$X_q(t) = X_q(kT_s), \qquad kT_s \leqslant t < (k+1)T_s$$

We see from Figure 10.11 that the quantized signal is a good approximation to the original signal. The quality of the approximation may be improved by a careful choice of x_i's and m_i's such that some measure of performance is optimized. The measure of performance that is most commonly used for evaluating the performance of a quantizing scheme is the output signal to quantizing noise power ratio defined as

$$\frac{S_q}{N_q}=\frac{E\{[X_q(kT_s)]^2\}}{E\{[X(kT_s)-X_q(kT_s)]^2\}} \tag{10.25}$$

Since the overall performance of digital transmission schemes for analog signals will be measured by signal-to-noise power ratios, and since the overall received signal quality will depend on the accuracy of representation of the sample values, the signal to quantizer noise power ratio, defined in Equation (10.25) is an appropriate measure of signal quality.

We will now consider several methods of quantizing the sampled values of a random process $X(t)$. For convenience, we will assume $X(t)$ to be a zero mean stationary random process with a pdf $f_X(x)$. We will use the abbreviated notation X to denote $X(kT_s)$ and X_q to denote $X_q(kT_s)$. The problem of quantizing consists of approximating the continuous random variable X by a discrete random variable X_q. We will use the mean squared error $E\{(X-X_q)^2\}$ as a measure of quantizing error.

10.3.1 Uniform Quantizing

In this method of quantizing, the range of the continuous random variable X is divided into Q intervals of equal length, say Δ. If the value of X falls in the ith *quantizing interval*, then the quantized value of X is taken to be the midpoint of the interval (see Figure 10.12). If a and b are the minimum and maximum values of X, respectively, then the *step size* or interval length Δ is given by

$$\Delta=(b-a)/Q \tag{10.26a}$$

The quantized output X_q is generated according to

$$X_q=m_i \quad \text{if} \quad x_{i-1}<X\leqslant x_i \tag{10.26b}$$

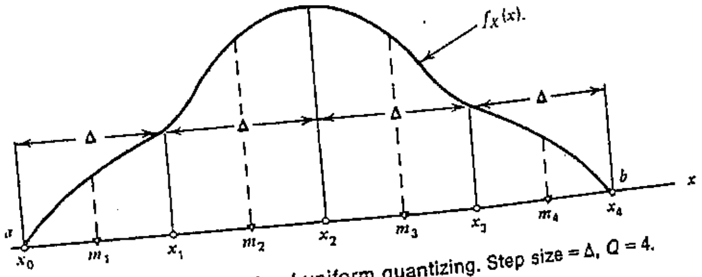


Figure 10.12 Example of uniform quantizing. Step size = Δ, $Q=4$.

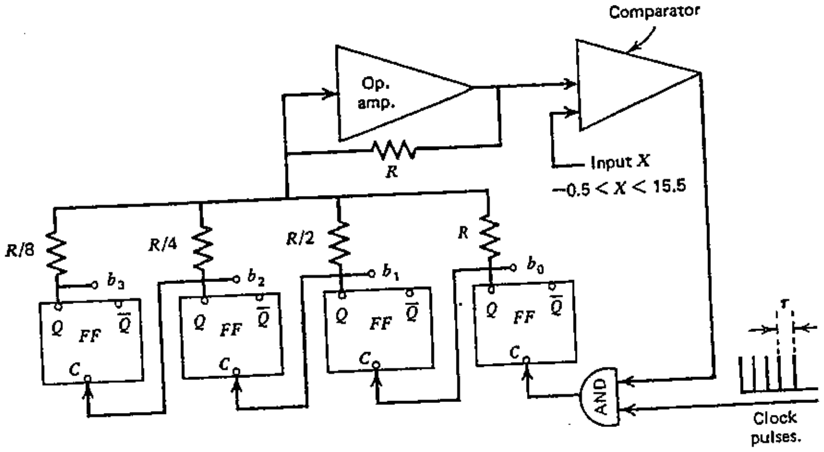


Figure 10.13 A sixteen level uniform quantizer. Step size = 1, and $b_3b_2b_1b_0$ gives the binary codeword for the level X_q. The clock rate is assumed to be much higher than the sampling rate and $Q=1$ volt.

where

$$x_i=a+i\Delta \tag{10.26c}$$

and

$$m_i=\frac{x_{i-1}+x_i}{2}, \quad i=1,2,\ldots,Q \tag{10.26d}$$

A uniform quantizer (A/D converter) that generates binary codes for the output levels is shown in Figure 10.13. It consists of a binary counter, a resistor matrix and summing device, and a comparator. The quantizer input X is assumed to be in the range −0.5 volt to 15.5 volts (if the range of X is outside this interval, then scaling and level shifting are necessary). When the quantizing is started, the counter is at zero and $X_q=0$. As the count is increased (while the AND gate is open) the value of X_q increases. As soon as X_q comes within $\frac{1}{2}$ volt of X, the comparator outputs a zero and closes the AND gate and blocks the clock pulses from incrementing the counter. The output of the operational amplifier represents the quantized value X_q of X, and the outputs of the flip-flops $b_3b_2b_1b_0$ provide a binary codeword for the output level. In this example, the numerical value of the binary codeword is equal to the value of X_q.

The quantizing noise power N_q for the uniform quantizer is given by

$$\begin{aligned} N_q &= E\{(X - X_q)^2\} \\ &= \int_a^b (x - x_q)^2 f_X(x)\,dx \\ &= \sum_{i=1}^{Q} \int_{x_{i-1}}^{x_i} (x - m_i)^2 f_X(x)\,dx \end{aligned} \tag{10.27a}$$

where $x_i = a + i\Delta$ and $m_i = a + i\Delta - \Delta/2$. The signal power S_q at the output of the quantizer can be obtained from

$$\begin{aligned} S_q &= E\{(X_q)^2\} \\ &= \sum_{i=1}^{Q} (m_i)^2 \int_{x_{i-1}}^{x_i} f_X(x)\,dx \end{aligned} \tag{10.27b}$$

The ratio S_q/N_q gives us a measure of fidelity of the uniform quantizer. This ratio can be computed if the pdf of X is known.

Example 10.1. The input to a Q-step uniform quantizer has a uniform pdf over the interval $[-a, a]$. Calculate the average signal to quantizer noise power ratio at the output.

Solution. From Equation (10.27a) we have

$$\begin{aligned} N_q &= \sum_{i=1}^{Q} \int_{x_{i-1}}^{x_i} (x - m_i)^2 \left(\frac{1}{2a}\right) dx \\ &= \sum_{i=1}^{Q} \int_{-a+(i-1)\Delta}^{-a+i\Delta} \left(x + a - i\Delta + \frac{\Delta}{2}\right)^2 \frac{1}{2a}\,dx \\ &= \sum_{i=1}^{Q} \left(\frac{1}{2a}\right)\left(\frac{\Delta^3}{12}\right) \\ &= \frac{Q\Delta^3}{(2a)12} = \frac{\Delta^2}{12}, \quad \text{since } Q\Delta = 2a. \end{aligned}$$

Now, the output signal power S_q can be obtained using Equation (10.27b) as

$$\begin{aligned} S_q &= \sum_{i=1}^{Q} (m_i)^2 \left(\frac{1}{2a}\right) \\ &= \frac{Q^2 - 1}{12} (\Delta)^2 \end{aligned}$$

and hence the average signal to quantizer noise power ratio is

$$\begin{aligned} \frac{S_q}{N_q} &= Q^2 - 1 \\ &\approx Q^2, \quad \text{when } Q \gg 1 \end{aligned} \tag{10.28a}$$

and

$$(S_q/N_q)_{\text{dB}} = 20 \log Q \tag{10.28b}$$

Equation (10.28) indicates that the fidelity of the quantizer increases with Q, the number of quantizer levels. If a large number of levels of small spacing are employed, then the output X_q can be made as near as desired to X, the input. The number of levels (Q) is determined by the desired transmission fidelity. It has been established experimentally that 8 or 16 levels are just sufficient to obtain good intelligibility of speech. But, the quantizer noise (whose power is more or less uniformly distributed throughout the signal band) can be easily heard in the background. For commercial use in standard voice telephony, a minimum of 128 levels are used to obtain a signal-to-noise ratio of 42 dB. This will require seven bits to represent each quantized sample and hence a larger transmission bandwidth than the unquantized analog voice signal.

The uniform quantizer yields the highest (optimum) average signal to quantizer noise power ratio at the output if the signal has a uniform pdf. The rms value of the quantizer noise is fixed at $\Delta/\sqrt{12}$ regardless of the value of the sample X being quantized. Hence if the signal $X(t)$ is small for extended periods of time, the apparent signal-to-noise ratio will be much lower than the design value. This effect will be particularly noticeable if the signal waveform has a large *crest factor* (the ratio of peak to rms value). For quantizing such signals it is advantageous to taper the spacing between quantizer levels with small spacings near zero and larger spacing at the extremes.

10.3.2 Nonuniform Quantizing

A nonuniform quantizer uses a *variable step size.* It has two important advantages over the uniform quantizer described in the preceding section. First, it yields a higher average signal to quantizing noise power ratio than the uniform quantizer when the signal pdf is nonuniform—which is the case in many practical situations. Secondly, the rms value of the quantizer noise power of a nonuniform quantizer is substantially proportional to the (instantaneous) sampled value X and hence the effect of quantizer noise is masked.

An example of nonuniform quantizing is shown in Figure 10.14. The input to the quantizer is a Gaussian random variable and the quantizer output is determined according to

$$\begin{aligned} X_q = m_i \quad &\text{if } x_{i-1} < X \leqslant x_i, \quad i = 1, 2, \ldots, Q \\ x_0 = -\infty, \quad &x_Q = \infty \end{aligned} \tag{10.29}$$

The step size $\Delta_i = x_i - x_{i-1}$ is variable. The quantizer end points x_i and the output

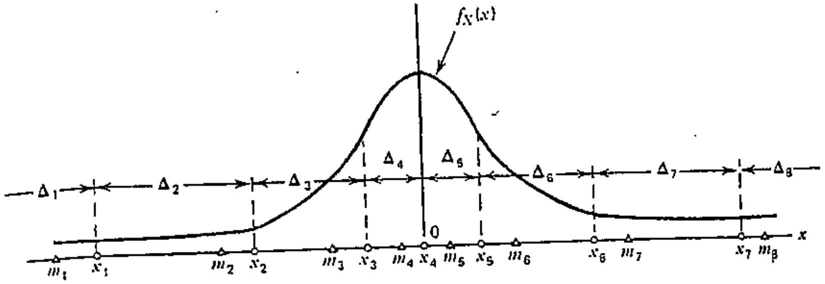


Figure 10.14 A nonuniform quantizer for a Gaussian variable. $x_0 = -\infty$, $x_Q = \infty$, $Q = 8$, and $\Delta_i = \Delta_{Q+1-i}$, $(i = 1, 2, 3, 4)$.

levels m_i are chosen to maximize the average signal to quantizing noise power ratio.

In practice, a nonuniform quantizer is realized by sample compression followed by a uniform quantizer. Compression transforms the input variable X to another variable Y using a nonlinear transformation

$$Y = g(X)$$

such that $f_Y(y)$ has a uniform pdf. Then, Y is uniformly quantized and transmitted (see Figure 10.15). At the receiver, a complementary expander with transfer characteristic g^{-1} restores the quantized values of X. The compresser and expander taken together constitute a *compander*. The most commonly used compander uses a logarithmic compression, $Y = \log X$, where the levels are crowded near the origin and spaced farther apart near the peak values of X.

Two commonly used logarithmic compression laws are the so-called μ and A compression laws defined by

$$|y| = \frac{\log(1 + \mu|x/x_{\max}|)}{\log(1 + \mu)}$$

and

$$|y| = \begin{cases} \dfrac{A|x/x_{\max}|}{1 + \log(A)}, & 0 \leq |x/x_{\max}| \leq 1/A \\[2ex] \dfrac{1 + \log(A|x/x_{\max}|)}{1 + \log(A)}, & 1/A \leq |x/x_{\max}| \leq 1 \end{cases}$$

Practical values of A and μ tend to be in the vicinity of 100. These two compression laws yield an average quantizing noise power that is largely independent of signal statistics (see Problems 10.11 and 10.12).

The design of an optimum nonuniform quantizer can be approached as follows. We are given a continuous random variable X with a pdf $f_X(x)$. We want to approximate X by a discrete random variable X_q according to Equation (10.29). The quantizing intervals and the levels are to be chosen such that S_q/N_q defined in Equation (10.24) is maximized. If the number of levels Q is large, then $S_q \approx E\{X^2\}$ and the ratio S_q/N_q is maximized when N_q is minimized. This minimizing can be done as follows. We start with

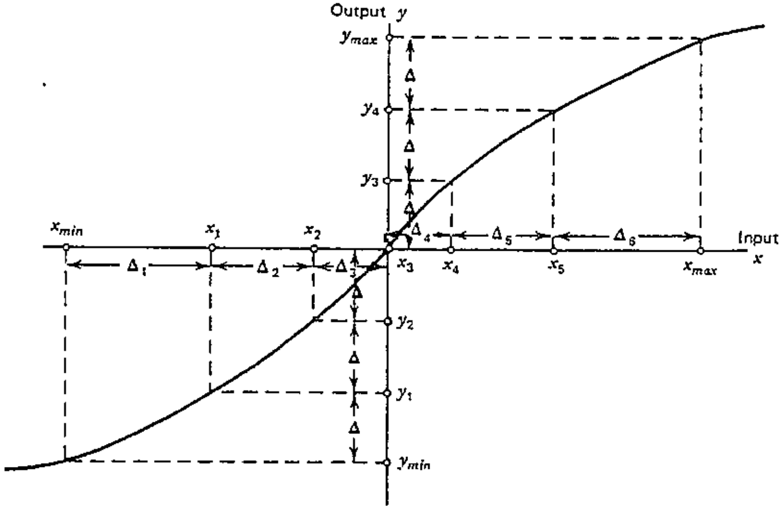


Figure 10.15 A compressor for converting a nonuniform quantizer to a uniform quantizer.

$$N_q = \sum_{i=1}^{Q} \int_{x_{i-1}}^{x_i} (x - m_i)^2 f_X(x)\,dx, \quad x_0 = -\infty \quad \text{and} \quad x_Q = \infty$$

Since we wish to minimize N_q for a fixed Q, we get the necessary* conditions by differentiating N_q with respect to x_i's and m_i's and setting the derivatives equal to zero:

$$\frac{\partial N_q}{\partial x_j} = (x_j - m_j)^2 f_X(x_j) - (x_j - m_{j+1})^2 f_X(x_j) = 0, \quad j = 1, 2, \ldots, Q - 1 \qquad (10.30a)$$

$$\frac{\partial N_q}{\partial m_j} = -2 \int_{x_{j-1}}^{x_j} (x - m_j) f_X(x)\,dx = 0, \quad j = 1, 2, \ldots, Q \qquad (10.30b)$$

*After finding all the x_i's and m_i's that satisfy the necessary conditions, we may evaluate N_q at these points to find a set of x_i's and m_i's that yield the absolute minimum value of N_q. In most practical cases we will get a unique solution for Equations (10.30a) and (10.30b).

From Equation (10.30a) we obtain

$$x_j = \tfrac{1}{2}(m_j + m_{j+1})$$

or

$$m_j = 2x_{j-1} - m_{j-1}, \quad j = 2, 3, \ldots, Q \tag{10.31a}$$

Equation (10.30b) reduces to

$$\int_{x_{j-1}}^{x_j} (x - m_j) f_X(x)\, dx = 0, \quad j = 1, 2, \ldots, Q \tag{10.31b}$$

which implies that m_j is the centroid (or statistical mean) of the jth quantizer interval. The above set of simultaneous equations cannot be solved in closed form for an arbitrary $f_X(x)$. For a specific $f_X(x)$, a method of solving (10.30a) and (10.30b) is to pick m_1 and calculate the succeeding x_i's and m_i's using Equations (10.31a) and (10.31b). If m_1 is chosen correctly, then at the end of the iteration, m_Q will be the mean of the interval $[x_{Q-1}, \infty]$. If m_Q is not the centroid or the mean of the Qth interval, then a different choice of m_1 is made and the procedure is repeated until a suitable set of x_i's and m_i's is reached. The reader can write a computer program to iteratively solve for the quantizing intervals and the means.

The end points of the quantizer intervals and the output levels for a normal random variable have been computed by J. Max [1]. Attempts have also been made to determine the functional dependence of N_q on the number of levels Q. For a normal random variable with a variance of 1, Max [1] has found that N_q is related to Q by

$$N_q \approx 2.2Q^{-1.96}, \text{ when } Q \gg 1$$

If the variance is σ_X^2, then the preceding expression becomes

$$N_q \approx (2.2)\sigma_X^2 Q^{-1.96} \tag{10.32}$$

Now, if we assume X to have zero mean, then $S_q \approx E\{X^2\} = \sigma_X^2$, and hence

$$S_q/N_q \approx (0.45)Q^{1.96} \tag{10.33}$$

Equation (10.33) can be used to determine the number of quantizer levels needed to achieve a given average signal to quantizer noise power ratio.

10.3.3 Differential Quantizing

In the preceding sections we saw that a continuous random process can be adequately represented by a sequence of its sampled values $\{X(kT_s)\}$ and that the individual samples $X(kT_s)$ can be approximated by a set of quantized levels. In the quantizing techniques we had considered thus far, each sample in the sequence $\{X(kT_s)\}$ was quantized independently of the value of the preceding sample. In many practical situations, due to the statistical nature of the message signal $X(t)$ and due to oversampling, the sequence $\{X(kT_s)\}$ will consist of samples that are correlated with each other. Differential quantizing schemes take into account the sample to sample correlation in the quantizing process. For a given number of levels per sample, differential quantizing schemes yield a lower value of quantizing noise power than direct quantizing schemes. Before we look at differential quantizing schemes, let us consider the following example that illustrates the main advantage of differential quantizing schemes.

Example 10.2. The message signal $X(t)$ in a communication system is a zero mean stationary Gaussian random process, and it is sampled at a rate of 10,000 samples per second ($T_s = 0.1$ msec). The normalized autocorrelation function $R_{XX}(\tau)/R_{XX}(0)$ has a value of 0.8 when $\tau = 0.1$ msec. Two quantizing schemes being considered are:

(a) a nonuniform quantizer with $Q = 32$, operating on each sample independently,
(b) a differential quantizer with $Q = 32$ which operates on successive differences $\{X(kT_s) - X((k-1)T_s)\}$.

Assuming that the mean squared error due to quantizing a normal random variable with variance σ^2 is $2\sigma^2 Q^{-2}$, find the mean squared error of the quantizing schemes given above.

Solution.
(a) For the quantizer operating independently on each sample, the mean squared error is given by

$$N_q = 2\sigma_X^2 Q^{-2} \approx (2)(10^{-3})\sigma_X^2$$

(b) In the differential quantizing scheme, the variable being quantized is

$$Y = X(kT_s) - X[(k-1)T_s]$$

and the variance of Y is given by

$$\begin{aligned} \sigma_Y^2 &= \sigma_X^2(kT_s) + \sigma_X^2[(k-1)T_s] - 2E\{X(kT_s)X[(k-1)T_s]\} \\ &= 2\sigma_X^2\left(1 - \frac{R_{XX}(T_s)}{R_{XX}(0)}\right) \\ &= 0.4\sigma_X^2 \end{aligned}$$

Hence the mean squared error of the differential quantizing is given by

$$\begin{aligned} N_q &= 2(0.4)\sigma_X^2 Q^{-2} \\ &= 0.8\sigma_X^2(Q^{-2}) \approx (0.8)(10^{-3})\sigma_X^2 \end{aligned}$$

which is considerably less than the error associated with the direct quantizing scheme.

The preceding example illustrates that differential quantizing yields a lower mean squared error than direct quantizing if the samples are highly correlated. This error reduction is always possible as long as the sample to sample correlation is nonzero. The largest error reduction occurs when the differential quantizer operates on the difference between $X(kT_s)$ and the minimum mean squared error estimator $\hat{X}(kT_s)$ of $X(kT_s)$. Such a quantizer using a linear minimum mean squared estimator* of $X(kT_s)$ based on the quantized values of preceding samples is shown in Figure 10.16. The difference $Y(kT_s) = X(kT_s) - \hat{X}(kT_s)$ is quantized and transmitted. Both the transmitter and the receiver use a predictor of the form

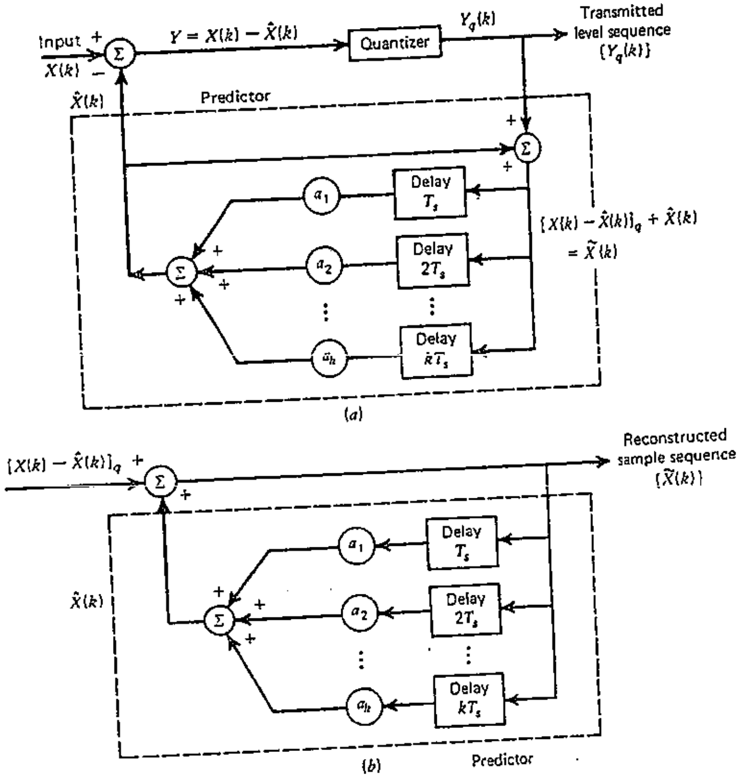


Figure 10.16 Differential quantizing scheme. (a) Differential quantizer at the transmitter. (b) Sample reconstruction scheme at the receiver.

$$\hat{X}(kT_s) = a_1\tilde{X}[(k-1)T_s] + a_2\tilde{X}[(k-2)T_s] + \cdots + a_n\tilde{X}[(k-n)T_s]$$

where

$$\tilde{X}(kT_s) = \hat{X}(kT_s) + [X(kT_s) - \hat{X}(kT_s)]_q$$

In the preceding equation $\tilde{X}$ denotes the reconstructed value of X and the subscript q denotes quantized values. The coefficients $a_1, a_2, \ldots, a_n$ are chosen such that $E\{[X(kT_s) - \hat{X}(kT_s)]^2\}$ is minimized. The differential quantizer discussed in Example 10.2 uses a predictor of the form $\hat{X}(kT_s) = X[(k-1)T_s]$.

The mean squared error of a differential quantizing scheme will be proportional to $E\{[X(kT_s) - \hat{X}(kT_s)]^2\}$, whereas the mean squared error of a direct quantizing scheme will be proportional to $E\{[X(kT_s)]^2\}$. If the predictor is good, which will be the case if the samples are highly correlated, then the mean squared error of the differential quantizer will be quite small. However, it must be pointed out here that the differential quantizer requires more hardware.

10.4 CODED TRANSMISSION OF ANALOG SIGNALS

After the output of an analog information source is sampled and quantized, the sequence of output levels $\{X_q(kT_s)\}$ can be transmitted directly using a Q-ary PAM. Alternatively, we may represent each quantized level by a code number and transmit the code number rather than the sample value itself. Source coding techniques discussed in Chapter 4 could be used to arrive at an optimum way of representing levels by code words. This system of transmission in which sampled and quantized values of an analog signal are transmitted via a sequence of codewords is called *Pulse Code Modulation* (PCM).

The important features of PCM are shown in Figure 10.17 and Table 10.1. We assume that an analog signal $x(t)$ with $\max|x(t)| < 4$ volts is sampled at the rate of r_s samples per second. The sampled values are quantized using a uniform quantizing rule with 16 steps ($Q = 16$) of equal step size $\Delta = 0.5$ volt. The quantizer end points are $-4, -3.5, -3, \ldots, 3.5, 4$ volts and the output levels are $-3.75, -3.25, \ldots, 3.25$, and 3.75 volts. Table 10.1 shows a sequence of sample values and the corresponding quantized levels. The 16 output levels are arbitrarily assigned level numbers $0, 1, 2, \ldots, 15$. These level numbers are

*For a good discussion of minimum mean squared error estimation, see Papoulis [2].

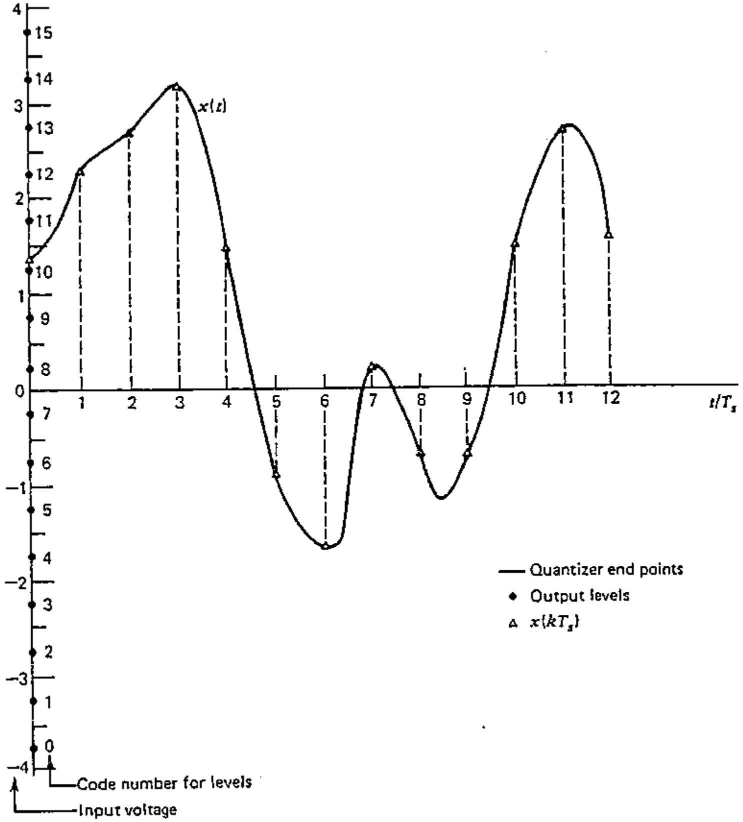


Figure 10.17 PCM example. Coded representation is given in Table 10.1.

shown encoded in binary and quarternary form in Table 10.1. The binary code is the binary representation of the level numbers. The quarternary code is easily derived from the binary code by segmenting each 4-bit binary word into two 2-bit binary words, and then converting each group of two binary digits to real integers. Now, if we are transmitting the sampled analog signal directly using analog PAM, we would transmit the sampled values 1.3, 2.3, 2.7, ... The symbol rate will be equal to the sampling rate r_s. If we are transmitting the quantized sample values using a 16-level discrete PAM, we would transmit the quantized levels 1.25, 2.25, 2.75, . . . , at a rate of r_s levels per second. In binary PCM, we would transmit the bit sequence 101011001101 . . . , at a bit rate of $4r_s$ bits/sec. Finally, if we use quarternary PCM, we will transmit the digit sequence 22303132 . . . , at a rate of $2r_s$ digits/sec. Each digit in this sequence can have one of four values.

Table 10.1. Quantizing and coding of an analog signal

Sampled values of an analog signal	1.3	2.3	2.7	3.2	1.1	−1.2	−1.6	0.1	−1.2
Nearest quantizer level	1.25	2.25	2.75	3.25	1.25	−1.25	−1.75	0.25	−1.25
Level number	10	12	13	14	10	5	4	8	5
Binary code	1010	1100	1101	1110	1010	0101	0100	1000	0101
Quarternary code	22	30	31	32	22	11	10	20	11

Several versions of PCM schemes are currently being used; two most commonly used versions are the differential pulse code modulation (DPCM) schemes and the delta modulation (DM) schemes. DPCM systems use differential quantizers and PCM encoders. DM schemes use a differential quantizer with *two* output levels Δ or −Δ; these two levels are encoded using a single binary digit before transmission. Thus, DM is a special case of DPCM.

In the following sections, we will discuss PCM, DPCM, and DM schemes in detail, and derive expressions for signal-to-noise ratios at the output of the receivers. Finally, we will compare the performance of these coded transmission schemes with the performance of analog modulation schemes such as AM and FM.

10.4.1 The PCM System

A PCM communication system is shown in Figure 10.18. The analog signal $X(t)$ is sampled and then the samples are quantized and encoded. For the purposes of analysis and discussion we will assume that the encoder output is a binary sequence. In the example shown in Figure 10.17 the binary code has a numerical significance that is the same as the order assigned to the quantized levels. However, this feature is not essential. We could have used arbitrary ordering and codeword assignment as long as the receiver knows the quantized sample value associated with each code word.

The combination of the quantizer and encoder is often called an *analog to*

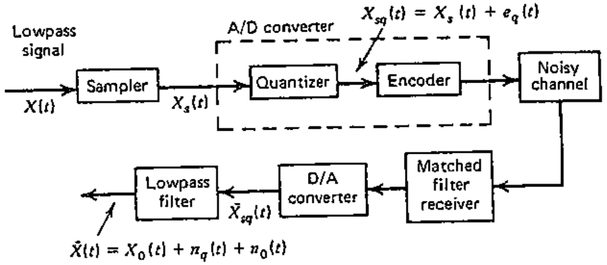


Figure 10.18 Block diagram of a PCM system. $n_q(t)$ is the noise due to quantizing and $n_0(t)$ is the noise due to bit errors caused by channel noise.

digital (A to D or A/D) *converter*. The sampler in practical systems is usually a *sample and hold* device. The combination of the sample and hold device and the A/D converter accepts analog signals and replaces it with a sequence of code symbols. A more detailed diagram of this combination, sometimes called a *digitizer*, is shown in Figure 10.19.

The digitally encoded signal is transmitted over the communication channel to the receiver (shown in Figure 10.20). When the noisy version of this signal arrives at the receiver, the first operation performed is the separation of the signal from the noise. Such separation is possible because of the quantization of the signal. A feature that eases this task of separating signal and noise is that during each bit interval the receiver (matched filter) has only to make the simple decision of whether a 0 or a 1 has been received. The relative reliability of this decision in binary PCM over the multivalued decision required for direct Q-ary PAM is an important advantage for binary PCM.

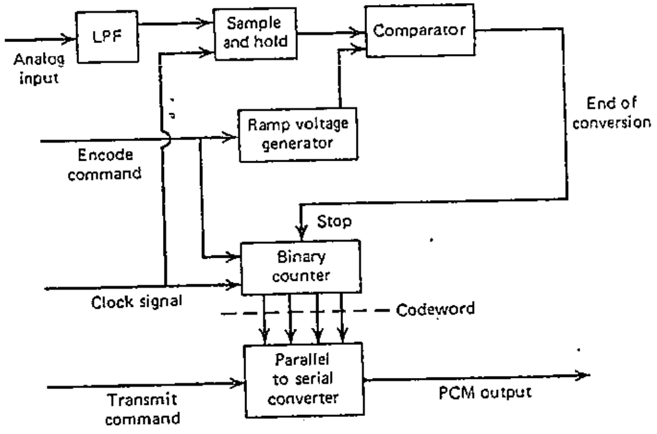


Figure 10.19 Elements of a PCM transmitter.

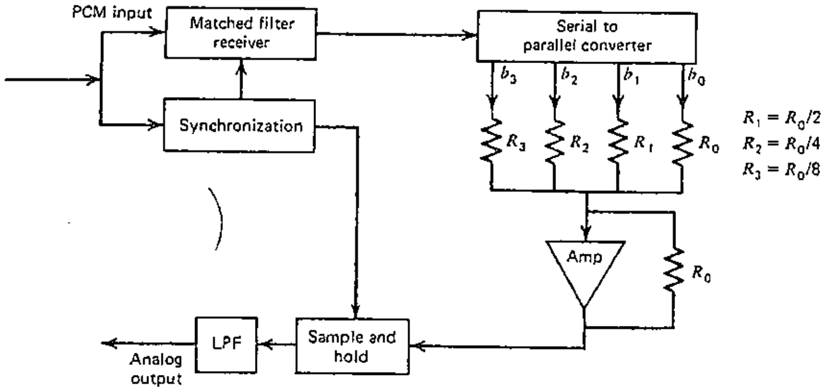


Figure 10.20 Elements of a PCM receiver.

After decoding a group of binary digits representing the codeword for the quantized value of the sample, the receiver has to assign a signal level to the codeword. The functional block that performs this task of accepting sequences of binary digits and generating appropriate sequences of levels is called a *digital to analog* (D/A) *converter*. The sequence of levels that appear at the output of the D/A converter as a Q-level PAM waveform is then filtered to reject any frequency components lying outside of the baseband. The reconstructed signal $\tilde{X}(t)$ is identical with the input $X(t)$ except for the quantization noise $n_q(t)$ and another noise component $n_0(t)$ that results from decoding errors due to the channel noise.

Figures 10.18–10.20 do not show signal companding components, and timing recovery networks.

10.4.2 Bandwidth Requirements of PCM

Since PCM requires the transmission of several digits for each message sample, it is apparent that the PCM bandwidth will be much greater than the message bandwidth. A lower bound on the bandwidth can be obtained as follows. If the message bandwidth is f_x, then the quantized samples occur at a rate $f_s(\geq 2f_x)$ samples per second. If the PCM system uses M channel symbols

(M-ary transmission) to represent the Q quantizer levels, then each codeword would consist of γ digits, where

$$\gamma = \log_M(Q), \quad M \leqslant Q$$

since there are M^γ different possible codewords and $M^\gamma \geqslant Q$ for unique coding. Thus the channel symbol rate is $r_s = \gamma f_s \geqslant 2\gamma f_x$. Recalling that for discrete baseband PAM signalling we need a bandwidth $\geqslant r_s/2$ Hz, we obtain the bandwidth of the PCM signal as

$$B_{\text{PCM}} \geqslant \gamma f_x \tag{10.34}$$

For binary PCM, the bandwidth required is greater than or equal to $f_x \log_2 Q$.

As an illustration of the bandwidth requirements let us consider the digital transmission of telephone-quality voice signal. While the average voice spectrum exceeds well beyond 10 kHz, most of the energy is concentrated in the range 100 to 600 Hz and a bandwidth of 3 kHz is sufficient for intelligibility. As a standard for telephone systems, the voice signal is first passed through a 3 kHz lowpass filter and then sampled at $f_s = 8000$ samples per second. Each sample is then quantized into one of 128 levels. If these samples are transmitted using binary PCM, then the bandwidth required will be larger than $(8000)(\frac{1}{2})(\log_2 128) = 28$ kHz, which is considerably greater than the 3 kHz bandwidth of the voice signal.

10.4.3 Noise in PCM Systems

It is shown in Figure 10.18 that the output $\tilde{X}(t)$ in a PCM system can be written as

$$\tilde{X}(t) = X_0(t) + n_q(t) + n_0(t) \tag{10.35}$$

where $X_0(t) = kX(t)$ is the signal component in the output; $n_q(t)$ and $n_0(t)$ are two noise components. The first noise waveform $n_q(t)$ is due to quantization and the additional noise waveform $n_0(t)$ is due to the additive channel noise. The overall signal-to-noise ratio at the baseband output, which is used as a measure of signal quality, is defined as

$$\left(\frac{S}{N}\right)_0 = \frac{E\{[X_0(t)]^2\}}{E\{[n_q(t)]^2\} + E\{[n_0(t)]^2\}} \tag{10.36}$$

The average noise power at the output, $E\{[n_q(t)]^2\}$ and $E\{[n_0(t)]^2\}$, can be calculated as follows.

Quantization Noise in PCM Systems. If we assume that ideal impulse sampling is used in the PCM system, then the output of the sampler is

$$X_s(t) = X(t) \sum_{k=-\infty}^{\infty} \delta(t - kT_s)$$

The quantized signal $X_{sq}(t)$ can then be expressed as

$$\begin{aligned} X_{sq}(t) &= X_q(t) \sum_k \delta(t - kT_s) \\ &= X(t) \sum_k \delta(t - kT_s) + [X_q(t) - X(t)] \sum_k \delta(t - kT_s) \\ &= \sum_k [X(kT_s)\delta(t - kT_s) + e_q(kT_s)\delta(t - kT_s)] \end{aligned}$$

where $e_q(t)$ is the error introduced by the quantizing operation. Using the results derived in Chapter 3, we can obtain the power spectral density of e_q as

$$G_{e_q}(f) = \frac{1}{T_s} E\{e_q^2(kT_s)\} \tag{10.37}$$

assuming that $E[e_q(kT_s)] = 0$ and $E\{e_q(kT_s)e_q[(k+j)T_s]\} = 0$. The mean squared error due to quantizing, $E\{e_q^2(kT_s)\}$, will depend on the signal statistics and the method of quantizing. For comparison purposes, let us assume a uniform quantizer operating on $X(t)$ having a uniform pdf over the interval $[-a, a]$. Then we have

$$E\{e_q^2(kT_s)\} = \Delta^2/12$$

where Δ is the step size, and

$$G_{e_q}(f) = \frac{1}{T_s}\left(\frac{\Delta^2}{12}\right)$$

If we ignore the effects of channel noise temporarily, then the noise component $n_q(t)$ has a power spectral density

$$G_{n_q}(f) = G_{e_q}(f)|H_R(f)|^2$$

where $H_R(f)$ is the transfer function of the lowpass filter used for reconstructing the signal. Assuming $f_s = 2f_x$ and $H_R(f)$ to be an ideal lowpass filter with a bandwidth f_x, we have

$$G_{n_q}(f) = \begin{cases} G_{e_q}(f) & |f| < f_x \\ 0 & \text{elsewhere} \end{cases}$$

Hence,

$$E\{n_q^2(t)\} = \int_{-f_x}^{f_x} G_{n_q}(f)\, df = \frac{1}{T_s^2}\left(\frac{\Delta^2}{12}\right)$$

The output signal component $X_0(t)$ is the response of the lowpass filter to

$X(t)\sum_k \delta(t-kT_s)$. We can calculate $E\{[X_0(t)]^2\}$ as

$$E\{[X_0(t)]^2\} \approx \frac{Q^2}{T_s^2}\left(\frac{\Delta^2}{12}\right) \tag{10.38}$$

where Q is the number of quantizer levels. Thus, the average signal to quantizer noise power ratio at the output of the PCM system is given by

$$\frac{E\{[X_0(t)]^2\}}{E\{[n_q(t)]^2\}} = Q^2 \tag{10.39}$$

This result is the same as the normalized mean square error due to quantizing (Equation (10.28a)), that is,

$$\frac{E\{[X_q(kT_s)]^2\}}{E\{[X(kT_s)-X_q(kT_s)]^2\}} = Q^2$$

This coincidence is due to the assumption that ideal impulse sampling is used in the system.

Channel Noise in PCM Systems. Channel noise causes the matched filter detector to make an occasional error in decoding whether a binary 0 or 1 was transmitted. The probability of error depends on the type of signaling used and the average signal-to-noise power ratio at the receiver input.

Typically, binary PCM systems operate with small word sizes and low probabilities of error. Hence, the likelihood of more than a single bit error within a codeword can be ignored. As an example, if the bit error probability is $P_e = 10^{-4}$ and a word has eight bits, we may expect on the average one word error for every 1250 words transmitted.

The probability of more than one bit error per word in this example would be of the order of $\binom{8}{2}P_e^2$ or of the order of 10^{-7}. When a bit error occurs in a PCM system, the decoder incorrectly identifies the transmitted level, and the quantized value of the signal is thus incorrectly determined. The magnitude of the error will be small if the bit error occurred in the least significant bit position, and the error will be large if the bit error occurred in the most significant bit position within the codeword.

In order to calculate the effects of bit errors induced by channel noise, let us consider a PCM system using N-bit codewords ($Q = 2^N$). Let us further assume that a codeword used to identify a quantization level is in the order of numerical significance of the word, that is, we assign 00 . . . 00 to the most negative level, 00 . . . 01 the next level, and 111 . . . 11 to the most positive level. An error that occurs in the least significant bit of the codeword corresponds to an error in the quantized value of the sampled signal by amount Δ. An error in the next significant bit causes an error of 2Δ, and an error in the ith bit position causes an error of $(2^{i-1})\Delta$. Let us call the error Q_Δ.

Then, assuming that an error may occur with equal likelihood in any one of the N bits in the codeword, the variance of the error is

$$\begin{aligned} E\{Q_\Delta^2\} &= \frac{1}{N}\left(\sum_{i=1}^{N}(\Delta 2^{i-1})^2\right) \\ &= \frac{2^{2N}-1}{3N}\Delta^2 \approx \frac{2^{2N}}{3N}\Delta^2 \end{aligned}$$

for $N \geqslant 2$. The bit errors due to channel noise lead to incorrect values of $X_q(KT_s)$. Since we are treating $X_q(t)$ as an impulse sequence, these errors appear as impulses of random amplitude and of random times of occurrence. An error impulse occurs when a word is in error. The mean separation between bit errors is $1/P_e$ bits. Since there are N bits per codeword, the mean separation between words that are in error is $1/(NP_e)$ words, and the mean time between word errors is

$$T = T_s/(NP_e)$$

Using the results derived in Chapter 3, we can obtain the power spectral density of the thermal noise error impulse train as

$$\begin{aligned} G_{th}(f) &= \frac{1}{T}E\{Q_\Delta^2\} \\ &= \left(\frac{NP_e}{T_s}\right)\left(\frac{2^{2N}}{3N}\right)\Delta^2 \end{aligned}$$

At the output of the ideal lowpass filter, the thermal noise error impulse train produces an average noise power N_0 given by

$$N_0 = \int_{-f_x}^{f_x} G_{th}(f)\,df = \frac{2^{2N}\Delta^2 P_e}{3T_s^2} \tag{10.40}$$

Output *S/N* Ratio in PCM Systems. The performance of the PCM system, when used for transmitting analog signals, is measured in terms of the average signal-to-noise power ratio at the receiver output. Combining Equations (10.36), (10.38), (10.39), and (10.40), we have

$$\left(\frac{S}{N}\right)_0 = \frac{2^{2N}}{1+4P_e 2^{2N}} \tag{10.41}$$

In Equation (10.41), P_e denotes the probability of a bit error which depends on the method of transmission. For example, if PSK signaling scheme is used, we have

$$P_e = Q\left(\sqrt{\frac{2S_{av}T_b}{\eta}}\right) = Q\left(\sqrt{\frac{2S_{av}T_s}{\eta N}}\right)$$

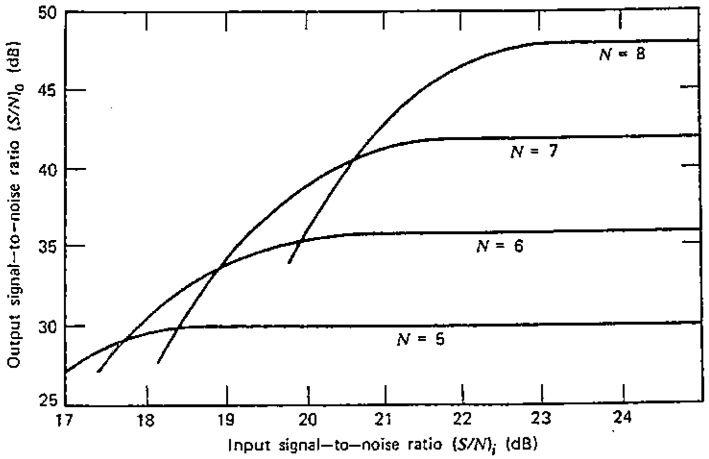


Figure 10.21 Output signal-to-noise ratio in PCM systems.

where S_{av} is the average signal power at the receiver input and T_b is the bit duration,

$$T_b = \frac{T_s}{N} = \frac{1}{2f_xN}$$

Hence,

$$P_e = Q\left(\sqrt{\frac{S_{av}}{\eta f_x N}}\right) \tag{10.42}$$

Similar expressions can be derived for other transmission schemes.

Plots of average signal-to-noise power ratio at the receiver input $(S/N)_i$ (defined as $S_{av}/\eta f_x$) versus the signal-to-noise ratio at the output $(S/N)_0$ for a PCM–PSK system are shown in Figure 10.21.

Plots shown in Figure 10.21 clearly indicate that the PCM system exhibits a threshold effect. For large values of $(S/N)_i$, P_e is small and hence

$$1 + 4P_e2^{2N} \approx 1$$

and

$$(S/N)_0 = 2^{2N} \approx (6N)\text{ dB} \tag{10.43a}$$

When $(S/N)_i$ is small, then we have

$$\left(\frac{S}{N}\right)_0 \approx \frac{2^{2N}}{4P_e2^{2N}} \approx \frac{1}{4P_e} \tag{10.43b}$$

The threshold point is arbitrarily defined as the $(S/N)_i$ at which $(S/N)_0$ given in Equation (10.43b) falls 1 dB below the value given in Equation (10.43a). The onset of threshold in PCM will result in a sudden increase in the output noise power. As the input signal power is increased, the output signal-to-noise power ratio $(S/N)_0$ reaches a value $(6N)$ dB that is independent of the signal power. Thus, above threshold, increasing signal power yields no further improvement in the $(S/N)_0$. The limiting value of $(S/N)_0$ depends only on the number of quantizer levels.

10.4.4 Differential PCM Systems

So far, we have discussed PCM systems using a fairly straightforward digital code for the transmission of analog signals. Several variations of PCM systems have been developed in recent years. We briefly describe two such types of systems here. Both systems use a differential quantizing scheme.

These systems are particularly more efficient when the sampled message signal has high sample to sample correlation. For example, in the transmission of picture (video) information, appreciable portions of the signal describe background information containing very little tonal variations. In such situations, if we use PCM, the codewords describe the value of the average background level; if these tonal values do not change appreciably, then we are essentially transmitting repeated sample values. One way to improve the situation is to send only the digitally encoded differences between successive samples. Thus a picture that has been quantized to 256 levels (eight bits) may be transmitted with comparable fidelity using 4-bit differential encoding. This reduces the transmission bandwidth by a factor of 2. PCM systems using differential quantizing schemes are known as differential PCM (DPCM) systems.

A differential PCM system that is particularly simple to implement results when the difference signal is quantized into two levels. The output of the quantizer is represented by a single binary digit, which indicates the sign of the sample to sample difference. This PCM system is known as delta modulation (DM). Delta modulation systems have an advantage over M-ary PCM and M-ary DPCM systems in that the hardware required for modulation at the transmitter and demodulation at the receiver are much simpler.

10.4.5 Delta Modulation Systems

The functional block diagram of a delta modulation system is shown in Figure 10.22. At the transmitter, the sampled value $X(kT_s')$ of $X(t)$ is compared with a predicted value $\hat{X}(kT_s')$ and the difference $X(kT_s') - \hat{X}(kT_s')$ is quantized into

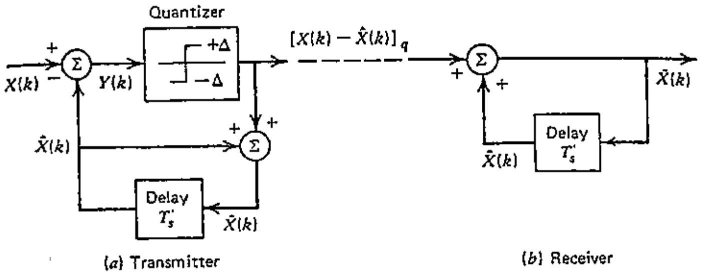


Figure 10.22 Discrete time model of a DM system. (a) Transmitter. (b) Receiver. Sampling rate $= f_s' = 1/T_s'$.

one of two values $+\Delta$ or $-\Delta$. The output of the quantizer is encoded using one binary digit per sample and sent to the receiver. At the receiver, the decoded value of the difference signal is added to the immediately preceding value of the receiver output. The operation of the delta modulation scheme shown in Figure 10.22 is described by the following equations:

$$\hat{X}(kT_s') = \tilde{X}((k-1)T_s') \tag{10.44}$$

where $\tilde{X}((k-1)T_s')$ is the receiver output at $t = (k-1)T_s'$ and

$$\tilde{X}(kT_s') = \hat{X}(kT_s') + [X(kT_s') - \hat{X}(kT_s')]_q = \tilde{X}((k-1)T_s') \pm \Delta \tag{10.45}$$

The delay element and the adder in Figures 10.22*a* and 10.22*b* can be replaced by an integrator whose input is an impulse sequence of period T_s' and strength $\pm\Delta$. This results in the system shown in Figure 10.23.

The operation of the delta modulation scheme shown in Figure 10.23 may be seen using the waveforms shown in Figure 10.24. The message signal $X(t)$ is compared with a stepwise approximation $\hat{X}(t)$ and the difference signal $Y(t) = X(t) - \hat{X}(t)$ is quantized into two levels $\pm\Delta$ depending on the sign of the difference. The output of the quantizer is sampled to produce

$$Y_{sq}(t) = \sum_{k=-\infty}^{\infty} \Delta \operatorname{sgn}[X(kT_s') - \hat{X}(kT_s')]\delta(t - kT_s') \tag{10.46}$$

The stepwise approximation $\hat{X}(t)$ is generated by passing the impulse waveform given in Equation (10.46) through an integrator that responds to an impulse with a step rise. Since there are only two possible impulse weights in $Y_{sq}(t)$, this signal can be transmitted using a binary waveform. The demodulator consists of an integrator and a lowpass filter.

In a practical delta modulation system, the lowpass filter in the receiver will, by itself, provide an approximate measure of integration. Hence we can eliminate the receiver integrator and depend on the filter for integration. At the transmitter, the sampling waveform $s_\delta(t)$ need not be an impulse waveform but a pulse waveform with a pulse duration that is short in comparison with the interval between pulses. Furthermore, the transmitter integrator need not be an ideal integrator—a simple RC lowpass filter will be adequate. These simplifications reduce the complexity of the hardware in DM systems considerably.

Some of the problems that occur when we use delta modulation to transmit an analog signal can be seen in the waveforms shown in Figure 10.24. Initially, let us assume that $\hat{X}(t) < X(t)$ so that the first impulse has a weight of Δ. When this impulse is fed back through the integrator, it produces a step change in $\hat{X}(t)$ of height Δ. This process continues through the *start up interval* until $\hat{X}(t)$ exceeds $X(t)$. During the start up interval the receiver output will differ considerably from the message signal $X(t)$. After the start up period, $\hat{X}(t)$ exhibits a *hunting* behavior when $X(t)$ remains constant. Hunting leads to idling noise. The sampling rate in a delta modulation scheme

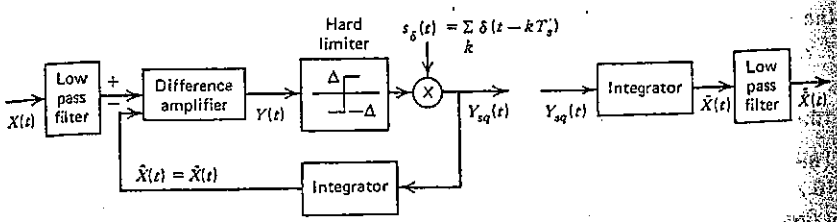


Figure 10.23 Hardware implementation of a DM system. (a) Modulator. (b) Demodulator.

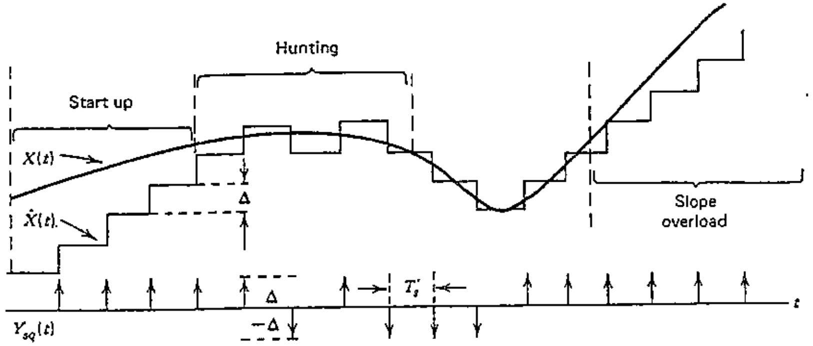


Figure 10.24 Delta modulation waveforms; $\hat{X}(t) = \tilde{X}(t)$.

will normally be much higher than the Nyquist rate and hence the rectangular idling noise waveform can be filtered or smoothed out by the receiver filter.

Slope Overloading. A serious problem in delta modulation schemes arises due to the rate of rise overloading. When $X(t)$ is changing, $\tilde{X}(t)$ and $\hat{X}(t)$ follow $X(t)$ in a stepwise fashion as long as successive samples of $X(t)$ do not differ by an amount greater than the step size Δ. When the difference is greater than Δ, $\hat{X}(t)$ and $\tilde{X}(t)$ can no longer follow $X(t)$. This type of overload is not determined by the amplitude of the message signal $X(t)$ but rather by its slope as illustrated in Figure 10.25; hence, the name slope overload.

To derive a condition for preventing slope overload in DM systems, let us assume that $X(t) = A\cos(2\pi f_x t)$. Then, the maximum signal slope is

$$\left[\frac{dX(t)}{dt}\right]_{\max} = A2\pi f_x$$

The maximum sample to sample change in the value of $X(t)$ then is $A2\pi f_x T'_s$. To avoid slope overload, this change has to be less than Δ, that is,

$$2\pi f_x T'_s A < \Delta$$

or, the peak signal amplitude at which slope overload occurs is given by

$$A = \frac{\Delta}{2\pi}\frac{f'_s}{f_x} \tag{10.47}$$

where $f'_s = 1/T'_s$ is the sampling rate of the DM system. For a signal $X(t)$ with a continuous spectrum $G_X(f)$, we can still use Equation (10.47) to determine the point of slope overload if f_x is taken to be the frequency beyond which $G_X(f)$ falls off at a rate greater than $1/f^2$. It has been determined experimentally that delta modulation will transmit speech signals without noticeable slope overload provided that the signal amplitude does not exceed the maximum sinusoidal amplitude given in Equation (10.47) with $f_x = 800$ Hz.

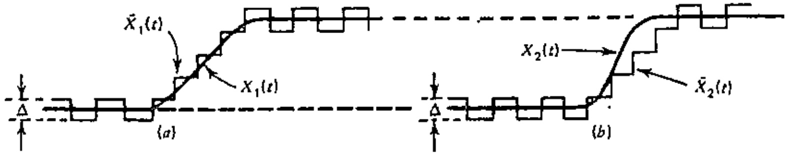


Figure 10.25 Slope overload in DM systems. Signals $X_1(t)$ and $X_2(t)$ have the same amplitude range. However, because of greater rate of rise, $X_2(t)$ causes a slope overload.

The problem of slope overloading in delta modulation systems can be alleviated by filtering the signal to limit the maximum rate of change or by increasing the step size and/or the sampling rate. Filtering the signal and increasing the step size will result in poor signal resolution, and increasing the sampling rate will lead to larger bandwidth requirements. A better way to avoid slope overload is to detect the overload condition and make the step size larger when overloading is detected. Systems using signal dependent step sizes are called adaptive delta modulation systems (ADM).

Adaptive Delta Modulation.* Hunting occurs in DM systems when the signal changes very slowly, and slope overloading occurs when the slope of the signal is very high. Both of these problems can be alleviated by adjusting the step size, in an adaptive fashion, in accordance with the signal being encountered. Ideally, the step size should be kept small when signal changes are small while increasing the step size in order to avoid slope overload when signal changes are large.

A DM system that adjusts its step size according to signal characteristics is shown in Figure 10.26. The step size is varied by controlling the gain of the integrator, which is assumed to have a low gain when the control voltage is zero and a larger gain with increasingly positive control voltage. The gain control circuit consists of an RC integrator and a square law device. When the

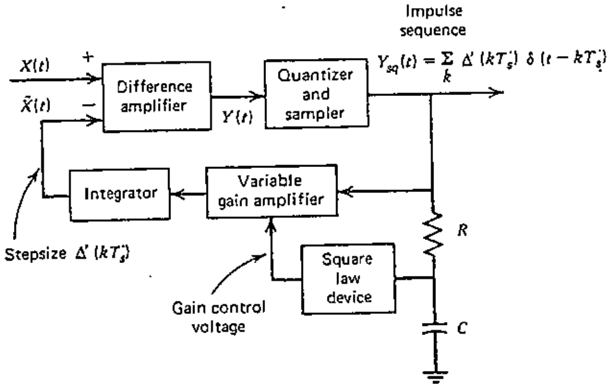


Figure 10.26 An adaptive delta modulator. The strength of the impulse $|\Delta'(kT'_s)|$ depends on the slope of the signal; the sign of $\Delta'(kT'_s)$ will be the same as the sign of $Y(kT'_s)$.

*Adaptive delta modulation is also known by the name continuous variable-slope delta modulation (CVSDM).

input signal is constant or slowly varying, the DM will be hunting and the modulator output will be a sequence of alternate polarity pulses. These pulses when integrated by the RC filter yield an average output of almost zero. The gain control input and hence the gain and the step size are small.

In case of a slope overload, the output of the quantizer will be a train of all positive or all negative pulses (see Figure 10.24). The integrator now provides a large control voltage and the gain of the amplifier is increased. Because of the squaring circuit, the amplifier gain will be increased no matter what the polarity of the pulses are. The net result is an increase in step size and a reduction in slope overload. The demodulator in an adaptive DM system will have an adaptive gain control circuit similar to the one used in the modulator.

10.4.6 Noise in Delta Modulation Systems

The output of the demodulator $\tilde{\tilde{X}}(t)$ differs from the input to the modulator $X(t)$ because of quantizing noise $n_q(t)$ and the noise due to transmission errors $n_0(t)$, that is,

$$\tilde{\tilde{X}}(t) = X_0(t) + n_q(t) + n_0(t) \tag{10.48}$$

where $X_0(t)$ is the output signal component (assumed to be equal to $X(t)$), and $n_0(t)$, $n_q(t)$ are the noise components at the output of the baseband filter. The overall signal quality in DM systems, as in PCM systems, is measured in terms of the signal-to-noise ratio at the output of the baseband filter. This ratio is defined as

$$\left(\frac{S}{N}\right)_0 = \frac{E\{[X_0(t)]^2\}}{E\{[n_q(t)]^2\} + E\{[n_0(t)]^2\}}$$

The average power content of the noise components can be calculated as follows.

Quantization Noise in DM Systems. To arrive at an estimate of the quantization noise power we write $X(t) = \tilde{X}(t) + e_q(t)$ where $|e_q(t)| = |X(t) - \tilde{X}(t)| \leq \Delta$ in the absence of slope overloading (Figure 10.24). The quantizing noise component $n_q(t)$ in Equation (10.48) is the response of the baseband filter to $e_q(t)$. If we assume a uniform pdf for $e_q(t)$, then

$$E\{[e_q(t)]^2\} = \int_{-\Delta}^{\Delta} \frac{1}{2\Delta} e^2\, de$$
$$= \Delta^2/3$$

It has been experimentally verified that the normalized power of the waveform $e_q(t)$ is uniformly distributed over the frequency interval $[0, f_s']$, where f_s' is the sampling rate. Thus, the power spectral density $G_{eq}(f)$ of $e_q(t)$ is given by

$$G_{eq}(f) = \begin{cases} \Delta^2/6f_s', & |f| < f_s' \\ 0, & \text{elsewhere} \end{cases}$$

Since $n_q(t)$ is the response of the baseband filter to $e_q(t)$, the normalized average power of the waveform $n_q(t)$ is given by

$$E\{[n_q(t)]^2\} = \int_{-f_x}^{f_x} G_{eq}(f)\, df$$
$$= \left(\frac{\Delta^2}{3}\right)\left(\frac{f_x}{f_s'}\right) \tag{10.49}$$

In order to compute the signal to quantizing noise power ratio, we need to calculate the signal power $E\{X_0^2(t)\}$. To simplify the calculation of signal power, let us take the worst case for delta modulation where all of the signal power is concentrated at the upper end of the baseband, that is, let us take

$$X(t) = A \cos 2\pi f_x t$$

Then,

$$X_0(t) = A \cos 2\pi f_x t$$

and

$$E\{X_0^2(t)\} = A^2/2 \tag{10.50a}$$

and to avoid slope overload we have, from Equation (10.47),

$$A = \frac{\Delta}{2\pi} \frac{f_s'}{f_x} \tag{10.50b}$$

Combining Equation (10.49) with (10.50), we obtain the output signal to quantizer noise power ratio as

$$\frac{E\{X_0^2(t)\}}{E\{n_q^2(t)\}} = \left(\frac{3}{8\pi^2}\right)\left(\frac{f_s'}{f_x}\right)^3 \tag{10.51}$$

We will see later on that the performance of DM systems as measured by signal to quantizer noise power ratio falls below the performance of PCM system using comparable bandwidth.

Channel Noise in DM Systems. When channel noise is present, the polarity of the transmitted waveform will be occasionally decoded incorrectly. Since the transmitted waveform is an impulse sequence of strength ±Δ, a sign error will result in an error impulse of strength 2Δ; the factor of 2 comes from the fact that an error reverses the polarity of the pulse. This channel-error noise appears at the receiver integrator input as a sequence of impulses with random times of occurrence and strength ±2Δ. The mean time of separation between these impulses is T_s'/P_e, where P_e is the bit error probability. The power spectral density of this impulse train can be shown to be white, with a magnitude of $4\Delta^2 P_e f_s'$. If we take the transfer function of the integrator to be $1/j\omega$, then the power spectral density of channel-error noise at the input to the baseband filter is given by

$$G_{th}(f) = \frac{4\Delta^2 P_e f_s'}{(2\pi f)^2} \tag{10.52}$$

It would appear now that to find the channel-error noise power at the output, $E\{n_0^2(t)\}$, all we need to do is to integrate $G_{th}(f)$ over the passband of the baseband filter. However, $G_{th}(f) \to \infty$ as $f \to 0$, and the integral of $G_{th}(f)$ over a range of frequencies including $f = 0$ is infinite. Fortunately, baseband filters have a low-frequency cutoff $f_1 > 0$; further, f_1 is usually very small compared to the high-frequency cutoff f_x. Hence

$$\begin{aligned} E\{n_0^2(t)\} &= 2\int_{f_1}^{f_x} G_{th}(f)\, df \\ &= \frac{2\Delta^2 P_e f_s'}{\pi^2}\left[\frac{1}{f_1} - \frac{1}{f_x}\right] \\ &\approx \frac{2\Delta^2 P_e f_s'}{\pi^2 f_1} \end{aligned} \tag{10.53}$$

since $f_1 \ll f_x$. Equation (10.53) shows that the output noise power due to bit errors depends on the low-frequency cutoff f_1 rather than the high-frequency cutoff f_x. Combining Equations (10.50), (10.51), and (10.53) we obtain the overall output signal-to-noise power ratio in a DM system as

$$\begin{aligned} \left(\frac{S}{N}\right)_0 &= \frac{E\{X_0^2(t)\}}{E\{n_q^2(t)\} + E\{n_0^2(t)\}} \\ &= \frac{(3f_s'^3/8\pi^2 f_x^3)}{1 + (6P_e f_s'^2/\pi^2 f_x f_1)} \end{aligned} \tag{10.54}$$

10.4.7 Comparison of PCM and DM Systems

We can now compare the performance of PCM and DM systems in terms of overall signal quality and equipment complexity. To ensure that the comparison is done under identical conditions, let us assume that both systems use approximately the same bandwidth for transmitting a baseband analog signal. If we use f_s and f_s' to denote the sampling rates of an N-bit PCM system and a DM system, then the bit rates for the systems are Nf_s and f_s', respectively. If the signal spectrum extends up to f_x Hz, then $f_s = 2f_x$ and identical bandwidth requirements imply that

$$f_s' = 2Nf_x$$

Signal-to-Noise Ratio. If the channel signal-to-noise ratio is high, then the performance of PCM and DM is limited by the quantization noise. The signal to quantizing noise power ratio for the PCM system is obtained from Equation (10.43),

$$(S_0/N_q)_{\text{PCM}} = Q^2 = 2^{2N}\ ; \quad N \geqslant 2$$

where $Q = 2^N$ is the number of quantizer levels. For the DM system, the corresponding ratio is given by Equation (10.51)

$$\begin{aligned} \left(\frac{S_0}{N_q}\right)_{\text{DM}} &= \frac{3}{8\pi^2}\left(\frac{f_s'}{f_x}\right)^3 \\ &\approx 0.3N^3 \end{aligned}$$

The preceding equations show that for a fixed bandwidth the performance of DM is always poorer than PCM. By way of an example, if the channel bandwidth is adequate for an 8-bit PCM code, then

$$(S_0/N_q)_{\text{PCM}} \approx 48 \text{ dB}$$

and

$$(S_0/N_q)_{\text{DM}} \approx 22 \text{ dB}$$

The performance of DM can be considerably improved by using a variable step size. Indeed, for speech transmission, it has been found that there is little difference in the performances of adaptive DM and PCM systems operating at a bit rate of about 64 kbits/sec.

The overall signal-to-noise ratio of a DM system is also lower than the overall S/N ratio of a PCM system using the same bandwidth. The extent of the difference in the signal quality depends on the characteristic of the signal. An example is given below:

Example 10.3. Compare the overall output S/N ratio for 8-bit PCM and DM systems used for transmitting a baseband signal whose spectrum is confined from 300 to 3000 Hz. Assume that both systems operate at a bit rate of 64 kbits/sec and use a PSK signaling scheme with $(S_{av}/\eta f_x) = 20$ dB.

Solution.

(a) PCM system. We have, $1/T_b = 64{,}000$, $(S_{av}/\eta f_x) = 100$, and

$$P_e = Q\left(\sqrt{\frac{2S_{av}T_b}{\eta}}\right) = Q(\sqrt{9.375}) \approx 10^{-3}$$

Hence,

$$\frac{S_0}{N_0} = \frac{2^{2N}}{1 + 4P_e 2^{2N}} \approx 24 \text{ dB}$$

(b) DM system. $f_1 = 300$, $f_x = 3000$, $f_s' = 64{,}000$, and

$$P_e = Q(\sqrt{9.375}) \approx 10^{-3}$$

From Equation (10.54) with $f_s' = 64{,}000$, $f_1 = 300$, and $f_x = 3000$, we have

$$S_0/N_0 \approx 20 \text{ dB}$$

Bandwidth Requirements. Since PCM and DM are now considered primarily for use in speech transmission, let us compare the BW requirements of these systems for speech transmission. With the use of PCM, speech transmission is found to be of good quality when $f_s = 8000$ and $N = 8$. The corresponding bit rate is 64 kbits/sec. To obtain comparable quality using delta modulation, the sampling rate has to be about 100 kbits/sec. However, it has been recently shown that with continuous variable slope delta (CVSD) modulation it is possible to achieve good signal quality at about 32 kbits/sec. It is reasonably accurate to conclude that PCM and (CVS) DM require approximately the same bandwidth for most analog signal transmission applications.

Equipment Complexity. The hardware required to implement DM is much simpler than that required for implementing PCM. Single integrated circuit chip (continuously-variable delta modulation) coder/decoders (called CODECS) are rapidly becoming available. In comparison, PCM coder/decoders require two chips for implementation: one chip for processing the analog signal and the second one to encode the sampled analog signal. Thus the PCM hardware is more expensive than the DM hardware.

10.4.8 *Q*-Level Differential PCM Systems

We conclude our treatment of digital transmission methods for analog signals with a brief description of a technique that combines the differential aspect of DM with the multilevel quantization of PCM. This technique, known as differential PCM (DPCM) or delta-PCM, uses a Q-level quantizer to quantize the difference signal $X(t) - \tilde{X}(t)$ (Figure 10.23). Thus the output of the sampler $Y_{sq}(t)$ is an impulse train in which the strength of the impulses can have one of Q possible values. (If $Q = 2$, then DPCM reduces to DM.) The value of the quantized error sample is represented by an N bit codeword $(Q = 2^N)$ and transmitted over the communication channel as a binary waveform.

The approximation $\tilde{X}(t)$ of $X(t)$ in a DPCM system has a variable step size ranging from $\pm\Delta$ to $\pm Q\Delta/2$, so $\tilde{X}(t)$ follows $X(t)$ more accurately. Thus, there will be much lower hunting noise, faster start-up, and less chance of slope overload, especially if a nonuniform quantizer is used.

The DPCM system combines the simplicity of DM and the multilevel quantizing feature of PCM. It has been found that, in many applications, the DPCM system with $Q \geq 4$ yields a higher signal to quantizer noise power ratio than ordinary PCM or DM using the same bit rate. In recent years DPCM systems have been used in the encoding and transmission of video signals (for example, in the Picturephone® system developed by AT&T). For broadcast quality black and white pictures, DPCM with $Q = 8 = 2^3$ gives acceptable video-signal reproduction, whereas straight PCM must have $Q = 256 = 2^8$ levels. Thus DPCM reduces transmission bandwidth by a factor of $\frac{3}{8}$. Comparable bandwidth reduction can be obtained for speech transmission also.

10.5 TIME-DIVISION MULTIPLEXING

Time-division multiplexing (TDM) is a technique used for transmitting several analog message signals over a communication channel by dividing the time frame into slots, one slot for each message signal. In comparison, frequency division multiplexing (FDM) divides the available bandwidth into slots, one slot for each message signal. The important features of TDM are illustrated in Figure 10.27.

Four input signals, all bandlimited to f_x by the input filters, are sequentially sampled at the transmitter by a rotary switch or *commutator*. The switch makes f_s revolutions per second and extracts one sample from each input during each revolution. The output of the switch is a PAM waveform containing samples of the input signals periodically interlaced in time. The samples from adjacent input message channels are separated by T_s/M, where

® Picturephone is a registered service mark of AT & T.

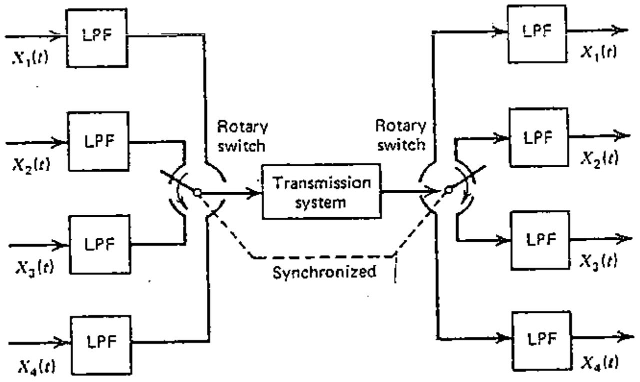


Figure 10.27 Block diagram of a four channel TDM system.

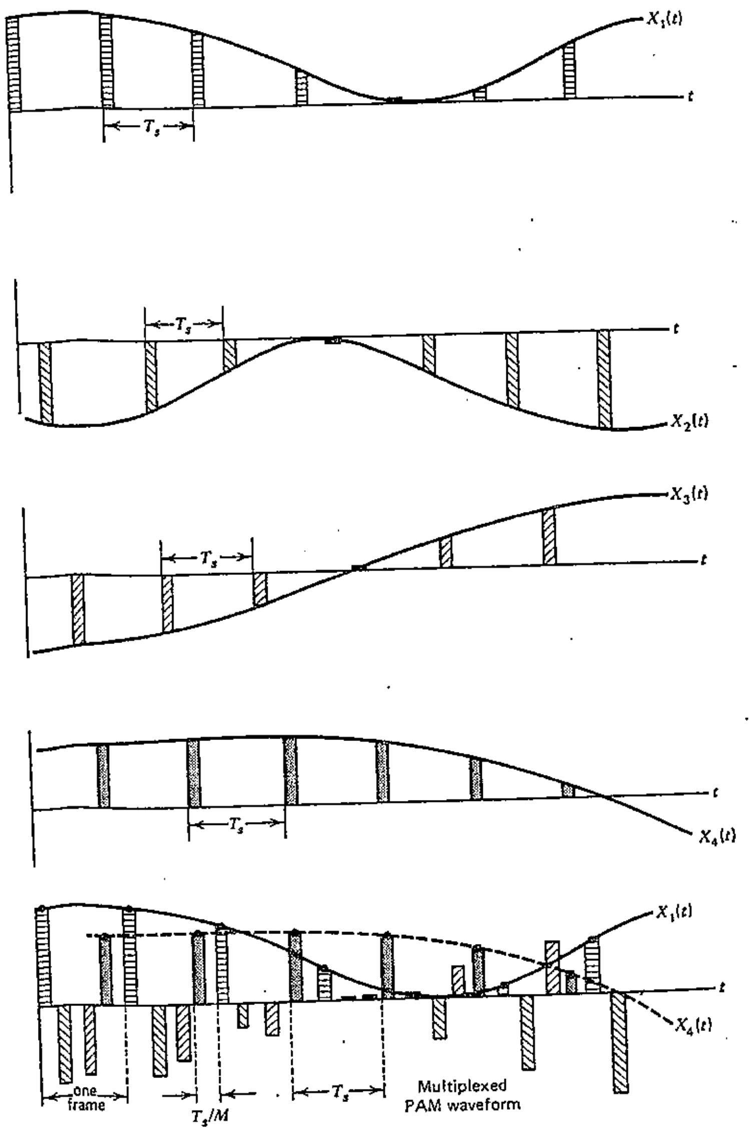


Figure 10.28 TDM waveforms; $T_s = 1/f_s$, where f_s is the number of revolutions per second of the rotary switch.

M is the number of input channels. A set of *M* pulses consisting of one sample from each of the *M*-input channels is called a *frame.* (See Figure 10.28.)

At the receiver, the samples from individual channels are separated and distributed by another rotary switch called a *distributor or decommutator.* The samples from each channel are filtered to reproduce the original message signal. The rotary switches at the transmitter and receiver are usually electronic circuits that are carefully synchronized. Synchronizing is perhaps the most critical aspect of TDM. There are two levels of synchronization in TDM: frame synchronization and sample (or word) synchronization. Frame synchronization is necessary to establish when each group of samples begin and word synchronization is necessary to properly separate the samples within each frame.

The interlaced sequence of samples may be transmitted by direct PAM or the sample values may be quantized and transmitted using PCM. Time-division multiplexed PCM is used in a variety of applications, the most important one is PCM telephone systems where voice and other signals are multiplexed and transmitted over a variety of transmission media including pairs of wires, wave guides, and optical fibers.

10.5.1 TDM–PCM Telephone System

The block diagram of a modular TDM–PCM telephone system designed by the American Telephone and Telegraph company is shown in Figure 10.29*a.* A 24-channel TDM multiplexer is used as the basic system, known as the T1 carrier system. Twenty-four voice signals are sampled at a rate of 8 kHz and the resulting samples are quantized and converted to 7-bit PCM codewords. At the end of each 7-bit codeword, an additional binary bit is added for synchronizing purposes. At the end of every group of twenty-four 8-bit codewords, another additional bit is inserted to give frame synchronization. The overall frame size in the T1-carrier is 193 bits, and the overall bit rate is 1.544 Mbits/sec. (See Figure 10.29*b.*)

The T1 system is designed primarily for short distance and heavy usage in metropolitan areas. The maximum length of the T1 system is now limited to 50 to 100 miles with a repeater spacing of 1 mile. The overall T-carrier system is made up of various combinations of lower order T-carrier subsystems designed for accommodating voice channels, Picturephone® service, TV signals, and (direct) digital data from data terminal equipment. A brief summary of the T-carrier TDM/PCM telephony system is given below in Table 10.2.

In addition to using metallic cable systems for transmission, optical fibers

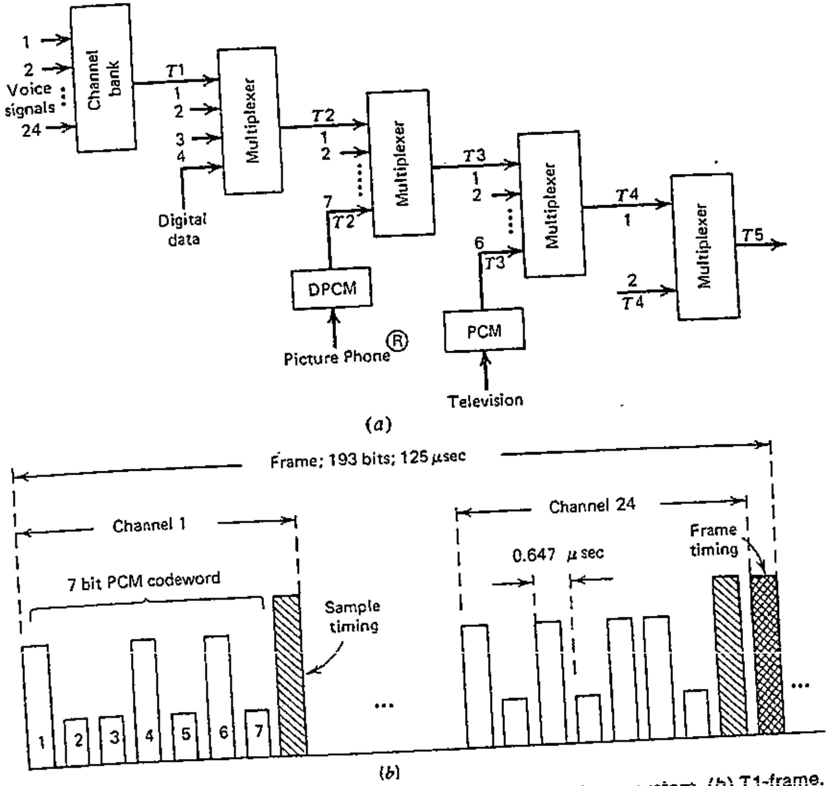


Figure 10.29 (a) The Bell system (AT & T) TDM/PCM telephony system. (b) T1-frame. Note that the timing pulses have larger amplitude than data pulses.

Table 10.2. T-Carrier Telephony System Specifications

System	Bit rate (Mbits/sec)	Medium	Repeater spacing'	Maximum length	System error rate
T1	1.544	wire pair	1 (mile)	50 (miles)	10^{-6}
T2	6.312	coax	2.5	500	10^{-7}
T3	44.736	coax	multiplexing only	—	—
T4	274.176	coax	1	500	10^{-6}
T5	560.160	coax	1	500	$(0.4)(10^{-8})$

with repeaters have been used to transmit binary data at speeds of 1.5, 3.6, 45, and 274 Mbits/sec corresponding to the speeds of the digital transmission hierarchy shown in Table 10.2 (see IEEE Spectrum, Feb., 1977).

10.5.2 Comparison of TDM and FDM

TDM and FDM techniques accomplish the same signal processing task. In TDM, the analog signals are separated in time but jumbled together in the frequency domain. In FDM the signals are separated in frequency domain but mixed together in time domain. From a theoretical point of view, the two systems may be viewed as dual techniques with neither one having any significant advantage over the other. However, from a practical viewpoint, TDM seems to be superior to FDM in at least two respects.

First, the TDM circuitry is much simpler than the FDM circuitry. FDM equipment consists of analog circuits for modulators, carrier generators, bandpass filters, and demodulators for *each* channel. In comparison, TDM circuitry is digital, consisting of a commutator and distributor. The digital circuitry is highly modular in nature and provides reliable and efficient operation.

A second advantage of TDM systems is the relatively small interchannel cross talk arising from nonlinearities in the circuits that handle the signals in the transmitter and the receiver. These nonlinearities produce intermodulation and harmonic distortion that affect both high-frequency and low-frequency channels in FDM systems. Thus the phase and amplitude linearity requirements of FDM circuits become very stringent when the number of channels being multiplexed is large. In contrast, there is no cross talk in TDM due to circuit nonlinearities if the pulses are completely isolated and nonoverlapping since signals from different channels are not handled simultaneously but are allotted different time intervals. Hence the linearity requirements of the TDM circuits are not quite as stringent as the FDM circuits. However, TDM cross talk immunity is contingent upon a wideband response and the absence of delay distortion. Disadvantages of TDM include the fact that pulse accuracy, timing jitter, and synchronization become major problems at high bit rates.

Finally, to complete our comparison of FDM and TDM, let us consider their bandwidth requirements. Let us assume that we have M input signals bandlimited to f_x Hz. With FDM using SSB modulation, the bandwidth of the multiplexed signal will be Mf_x. With TDM, if we assume a sampling rate of f_s for each channel, then the multiplexed signal consists of a series of sample points separated in time by $1/Mf_s$ sec (Figure 10.30*a*). By virtue of the sampling theorem, these points can be completely described by a continuous waveform $X_b(t)$ that is bandlimited to $Mf_s/2$ Hz. This waveform, even though

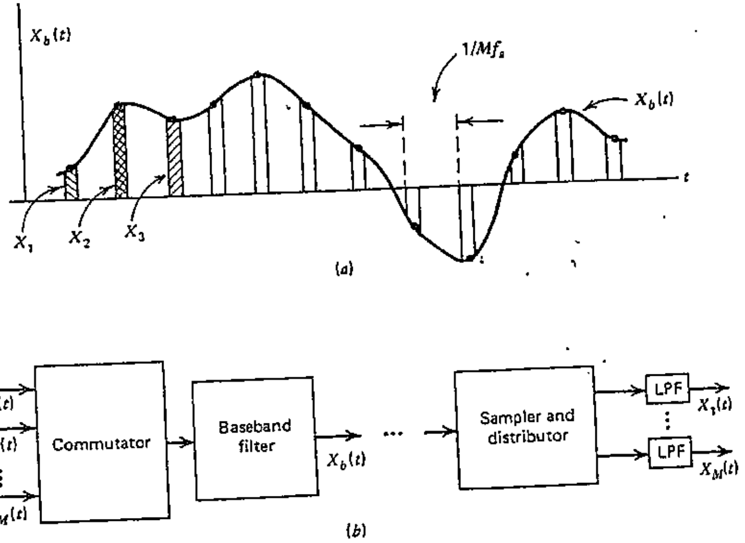


Figure 10.30 Baseband filtering of a TDM waveform.

it has no direct relation to the original messages, passes through the correct sample values at sampling times. $X_b(t)$ is obtained by lowpass filtering of the interleaved sample sequence. At the receiver, $X_b(t)$ is sampled and the sample values are distributed to appropriate channels by the distributor. If the sampling frequency is close to the Nyquist rate, that is, $f_s \approx 2f_x$, then the bandwidth of the filtered TDM waveform is Mf_x Hz, which is the same as the bandwidth of the FDM waveform.

10.5.3 Asynchronous TDM

In the preceding discussion of TDM systems we had assumed that the signals being multiplexed have comparable bandwidths and hence the sampling rate for each signal is the same. However, in many applications the signals to be time-division multiplexed have different bandwidths and hence they have to be sampled at different sampling frequencies. In these situations, we cannot multiplex these signals using the technique described previously, which employs a common clock rate for all the channels.

One method of combining a group of asynchronously sampled time-division multiplexed signals uses *elastic store* and *pulse stuffing*. An elastic storage device, which is essential for multiplexing asynchronous signals, stores a digital bit stream in such a manner that the bit stream may be read out at a rate different from the rate at which it was read in. One example of such a device is a tape recorder. Data can be recorded onto the tape and read out at a different rate by adjusting the tape speed during replay. Another example is a large (digital) buffer into which data can be read in at one rate and read out at a different rate.

To illustrate the use of elastic store and pulse stuffing in asynchronous TDM, consider the example of a satellite that records the results of a number of experiments and transmits them to the earth. Let us suppose that three experiments each lasting a duration of one second are performed simultaneously, and that their signals are sampled and stored in three separate digital storage devices. Let us assume that the three signals are sampled at rates 2000, 3000, and 5000 samples per second, respectively, and the samples are encoded using 8-bit PCM codewords. At the end of each 1-sec interval, the experiments are halted for one second during which time all of the data collected are transmitted to earth.

During transmission, each storage device can be emptied (played back) at the same rate (5000 samples per second), synchronously time-division multiplexed and a single TDM signal can be transmitted to earth. There is one major problem associated with this procedure. The first 2000 words of each signal can be multiplexed without any trouble. During the multiplexing of the next 2000 words there is no contribution from the first signal, and during the last 1000 words there is no contribution from the first and second signals. However, because of noise, the receiver will be reading words when no words from channels 1 and 2 are being transmitted. To avoid this erroneous interpretation of noise as signal, the time slots corresponding to signals that have already terminated are filled with dummy sequences of bits. These dummy sequences are carefully chosen and encoded so that the receiver recognizes them without difficulty. This technique is called pulse stuffing since it requires that digits or pulses be stuffed into spaces provided for the missing message bits. (See Figure 10.31.)

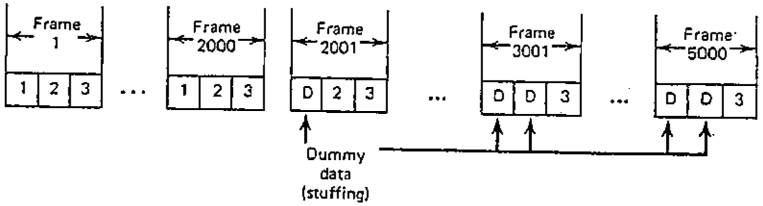


Figure 10.31 Example of pulse stuffing.

10.6 COMPARISON OF METHODS FOR ANALOG SIGNAL TRANSMISSION

In this section we provide a comparison of analog and digital methods for transmitting analog signals. First we will compare analog modulation methods with PCM methods in a qualitative manner. Later we will attempt to compare these methods in a quantitative manner in terms of signal-to-noise ratio at the receiver output and power bandwidth requirements using an information theoretic approach. In order to do this, we will define an ideal (but unrealizable) communication system using the Shannon–Hartley law. The performance of practical analog and digital modulation schemes will then be compared with the bounds set by the ideal system.

10.6.1 PCM versus Analog Modulation

PCM systems have certain inherent advantages over analog modulation schemes for transmitting analog signals. Some of these advantages are the following:

(1) In long distance communications, PCM signals can be completely regenerated at each repeater station if the repeater spacing is such that the magnitude of the noise is less than half the separation between levels (with a high probability). An example of signal regeneration at a repeater is shown in Figure 10.32. With the exception of occasional errors, a noise- and distortion-free signal is transmitted at each repeater. Further, the effect of noise does not accumulate and in designing repeaters one needs to be concerned only about the effects of channel noise between repeater stations. Repeaters for analog modulation schemes consist of amplifiers that raise the signal level at each transmitting station. While raising the signal level, the amplifier also raises the level of accompanying noise at each repeater station.

(2) At low input signal-to-noise ratios, the output signal-to-noise ratio of PCM systems is better than analog modulation schemes. A quantitative discussion of the effect of noise in various modulation schemes will be presented later on in this chapter.

(3) PCM systems can be designed to handle a variety of signals. A PCM system designed for analog message transmission can be readily adapted to handle other signals, particularly digital data.

(4) Modulation and demodulation circuitry in PCM systems are all digital thus affording high reliability and stability. Advances in integrated circuits have lowered the cost of these circuits considerably.

(5) It is easy to store and time scale PCM signals. For example, PCM data gathered intermittently from an orbiting satellite over a period of hours may be transmitted to the ground station in a matter of seconds when the satellite appears in the field of view of the receiving antenna. While this may also be accomplished using analog techniques, digital memories can perform the required storage very efficiently. Further, PCM signals can be time-division multiplexed easily.

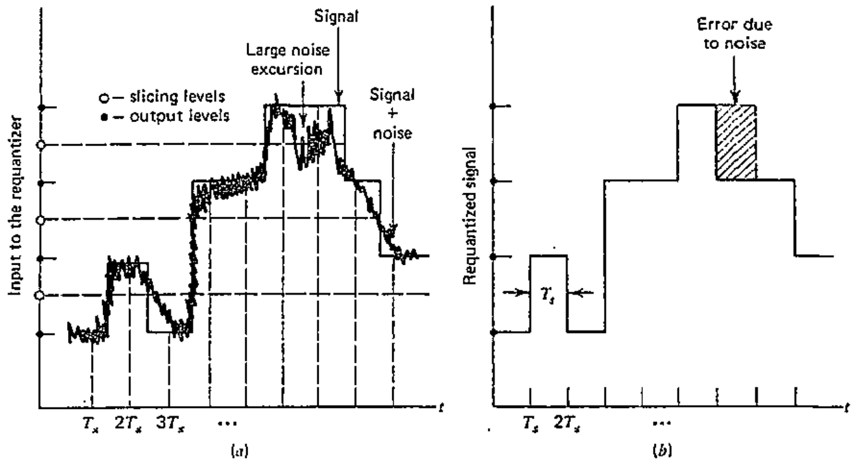


Figure 10.32. Signal regeneration by requantizing. (a) Noisy signal at the input of the repeater. (b) Signal at the repeater output; errors occur when the noise amplitude is large.

(6) With PCM systems, source coding and channel coding techniques can be used to reduce unnecessary repetition (redundancy) in messages and to reduce the effects of noise and interference. Coding may also be used to make the digital communication channel more secure.

(7) As we will see (in the next section) the exchange of bandwidth for power is easy to accomplish in PCM systems. Since PCM systems can be easily time scaled, time can also be exchanged for signal power. Thus the communication systems designer has added flexibility in the design of a PCM system to meet a given performance criteria.

Some of the advantages of PCM systems are offset by the fact that the complexity of a PCM system is greater than that required for other types of modulation systems. However, the complexity of a PCM system varies little as the number of message channels is increased. Hence, PCM systems can compare quite favorably with other systems when the number of channels is large.

10.6.2 Comparison of Communication Systems: Power-Bandwidth Exchange

In communication systems designed to handle analog message signals, the signal-to-noise ratios at various points in the systems are used to measure the signal quality at these points. Of particular interest are the signal-to-noise ratio at the input to the receiver and the signal-to-noise ratio at the receiver output. The signal-to-noise ratio at the input depends on the transmitted power and the ambient noise appearing at the receiver antenna. The output signal-to-noise ratio depends on the input signal-to-noise ratio and the type of modulation/demodulation processes used in the system. The ratio of the signal-to-noise ratio at the output and the input signal-to-noise ratio, called the detection gain (a measure of noise immunity), is widely used as a figure of merit for communication systems. We will use this figure of merit to compare the performance of several communication systems. We will first investigate the performance of an ideal (but unrealizable) communication system implied by the Shannon–Hartley law. We will then examine various practical communication systems to see how they measure up against the ideal system—particularly in the exchange of bandwidth for signal-to-noise ratio (or transmitted power).

An Ideal Communication System. Suppose that we have a communication system (Figure 10.33) for transmitting an analog message signal $X(t)$ bandlimited to f_x Hz. Further, suppose that an ideal system is available for this purpose and that the channel bandwidth is B_T and the noise power spectral density is $\eta/2$. Also, let us assume that the average signal power at the receiver is S_r and that the desired value of the output signal-to-noise ratio is $(S/N)_d$.

Now, the channel capacity of the system is given by the Shannon–Hartley law as

$$C = B_T \log_2[1+(S/N)_r] \tag{10.55}$$

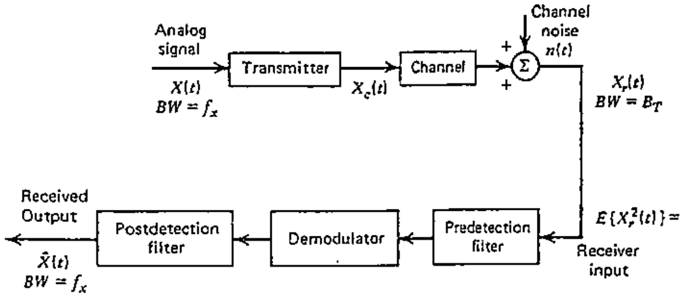


Figure 10.33 Block diagram of a communication system.

where $(S/N)_r$ is the signal-to-noise ratio at the receiver input. At the receiver output, the information rate can be no greater than

$$R_{\max} = f_x \log_2[1+(S/N)_d] \tag{10.56}$$

An optimum or ideal system is defined as one that is operating at its capacity, with maximum output rate. That is, for the ideal system, we have

$$R_{\max} = C$$

or

$$B_T \log_2[1+(S/N)_r] = f_x \log_2[1+(S/N)_d]$$

We can solve for $(S/N)_d$ at the output of the ideal system as

$$\begin{aligned}(S/N)_d &= [1+(S/N)_r]^{B_T/f_x} - 1\\ &\approx [1+(S/N)_r]^{B_T/f_x}\end{aligned} \tag{10.57}$$

when the signal-to-noise ratios are large. In Equation (10.57) the input signal-to-noise ratio $(S/N)_r$ is given by

$$\left(\frac{S}{N}\right)_r = \frac{S_r}{\eta B_T} \tag{10.58}$$

The ratio of transmission bandwidth B_T to message bandwidth f_x is called the *bandwidth expansion ratio* (or *bandwidth compression ratio* if the ratio is less than 1). If we let

$$\beta' = B_T/f_x$$

and

$$\alpha = \frac{S_r}{\eta f_x}$$

then we can rewrite Equation (10.57) as

$$\begin{aligned}\left(\frac{S}{N}\right)_d &= \left[1+\frac{\alpha}{\beta'}\right]^{\beta'} - 1\\ &\approx \left(\frac{\alpha}{\beta'}\right)^{\beta'}\end{aligned} \tag{10.59}$$

when the signal-to-noise ratios are large.

Equation (10.59) shows that, in an ideal system, the signal-to-noise ratio at the output and the bandwidth are *exponentially* related. This means that doubling the transmission bandwidth of an ideal system squares the output signal-to-noise ratio. Alternately, since $\alpha = S_r/\eta f_x$ is proportional to the transmitted power S_T, the transmitted power can be reduced to the square root of its original value without reducing $(S/N)_d$ by increasing the bandwidth by a factor of 2.

Example 10.4. Consider an ideal system, designed for transmitting an analog message signal, with the following parameters.

$$(S/N)_d = 60\ \text{dB}$$

Inband noise power $\eta f_x = 10^{-7}$ watt, and $f_x = 15$ kHz. Compare the power requirements (S_r) of the ideal system for the following transmission bandwidths: (a) $B_T = 15$ kHz. (b) $B_T = 75$ kHz. (c) $B_T = 5$ kHz.

Solution

(a) With $B_T = 15$ kHz, $\beta' = 1$: and with

$$(S/N)_d = 60\ \text{dB} = 10^6$$

we have (from Equation (10.59))

$$(\alpha/\beta')^{\beta'} \approx 10^6 \quad \text{or} \quad \alpha \approx 10^6$$

Since

$$\alpha = S_r/\eta f_x$$

we have

$$S_r = \alpha\eta f_x = (10^6)(10^{-7}) = 0.1\ \text{watt}$$
$$= 20\ \text{dBm}$$

(b) With $B_T = 75$ kHz, we have $\beta' = 5$ and

$$(\alpha/5)^5 \approx 10^6 \quad \text{or} \quad \alpha \approx 79.24$$

Hence,

$$S_r = \alpha\eta f_x = (79.24)(10^{-7})$$
$$\approx -21.02\ \text{dBm}$$

(c) With $B_T = 5$ kHz (bandwidth compression), the reader can verify that $\beta' = 0.333$ and

$$\alpha = (\tfrac{1}{3})(10)^{18}$$

or

$$S_r = (\tfrac{1}{3})(10)^{11}\ \text{watts (a colossal amount of power!)}$$

The preceding example illustrates that bandwidth expansion leads to a considerable reduction in power requirements. However, bandwidth compression leads to extremely large power requirements. We may generalize this conclusion and say that optimum bandwidth to power exchange is practical in one direction only, namely, the direction of increasing bandwidth and decreasing power.

Comparison of Communication Systems. We are now ready to compare the performance of existing communication systems with the ideal system discussed in the preceding section. When comparing existing systems with the ideal system we should remember the following points. The ideal system discussed in the preceding section was arrived at via an information-theoretic approach; the primary goal of the system was reliable information transfer in the sense of information theory. In systems designed for transmitting analog message signals it is very difficult to assess the information rate. Furthermore, the primary concern in such applications might be signal-to-noise ratios, threshold power, (no threshold effect in ideal systems!), and bandwidth requirements rather than channel capacity and its utilization.

A comparison of the performance of many practical systems with that of an ideal system is shown summarized in Table 10.3 and Figure 10.34. The results for SSB, AM, and DSB modulation are taken from Chapters 6 and 7. It is assumed that signal-to-noise ratios are large, all systems are above threshold, and that the message signal is normalized with $E\{X^2(t)\} = E\{\hat{X}^2(t)\} = \frac{1}{2}$. The result for the PCM system is obtained from Equation (10.41). The performance of the PCM system operating above threshold is limited by quantizing noise, and

$$(S/N)_d = Q^2$$

where Q is the number of quantizer levels. Now if we use an M-ary PCM, and if the sampling rate is $f_s = 2f_x$, then the transmission bandwidth B_T is $r_s/2$, where r_s is the channel symbol rate given by

$$r_s = (2f_x)\log_M Q$$

Table 10.3. Performance of communication systems. $\beta' = B_T/f_x$; $\alpha = S_r/\eta f_x$.

System	Bandwidth expansion	$(S/N)_d$
Ideal	β'	$(\alpha/\beta')^{\beta'}$
SSB Baseband	$\beta' = 1$	α
DSB	$\beta' = 2$	α
AM	$\beta' = 2$	$\alpha/3$
WBFM	$\beta' > 1$	$\frac{3}{8}\alpha\beta'^2$
PCM	$\beta' = \log_M(Q)$	$M^{2\beta'}$

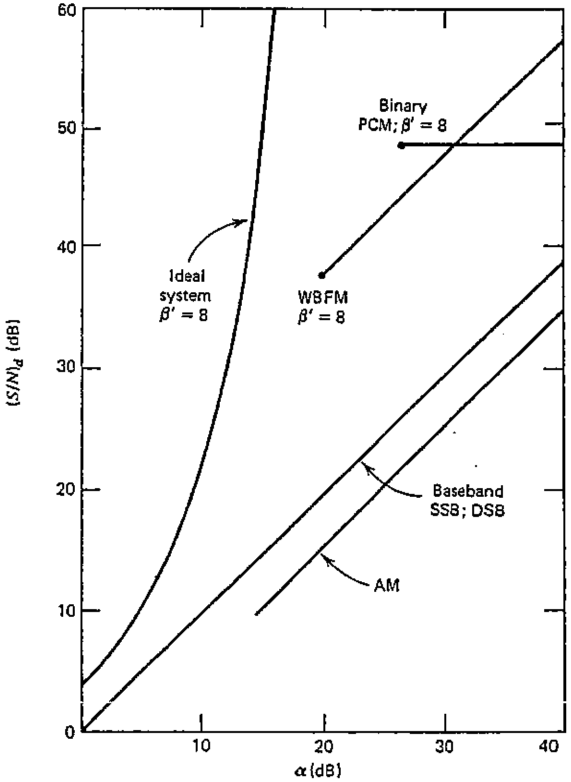


Figure 10.34 Signal-to-noise ratios in communication systems. $\alpha = (S/N)_r$; $\beta' = B_T/f_x$.

Hence

$$B_T = f_x \log_M Q$$

or

$$Q = M^{(B_T/f_x)} = M^{\beta'}$$

Thus the output signal-to-noise ratio for the PCM above threshold is

$$(S/N)_d = Q^2 = M^{2\beta'} \tag{10.60}$$

The results shown in Table 10.3 and Figure 10.34 indicate that none of the practical systems can match the signal-to-noise ratio improvement that is possible with an ideal system. This is due to the fact that practical systems, with the exception of PCM systems, do not have an exponential power-bandwidth dependence like the ideal system. While PCM does have an exponential power-bandwidth relationship, increasing transmitted power beyond threshold yields no further improvement in $(S/N)_d$ since the limiting value of $(S/N)_d$ is determined by quantization.

The SSB system offers as good a performance as an ideal system with $\beta' = 1$. However, the bandwidth ratio for SSB is fixed at $\beta' = 1$ and there is no possibility of trading bandwidth for noise reduction. AM and DSB systems with $\beta' = 2$ do not perform as well as an ideal system with $\beta' = 2$. Furthermore, like in SSB modulation, there are no possibilities of wideband noise reduction in AM and DSB systems.

Wideband FM systems offer good possibilities for noise reduction. But the performance of WBFM, like PCM and AM systems (using noncoherent demodulation procedures), falls short of the performance of ideal systems because of threshold effects, especially at low input signal-to-noise ratios.

Figure 10.35 shows the minimum values of input signal-to-noise ratios

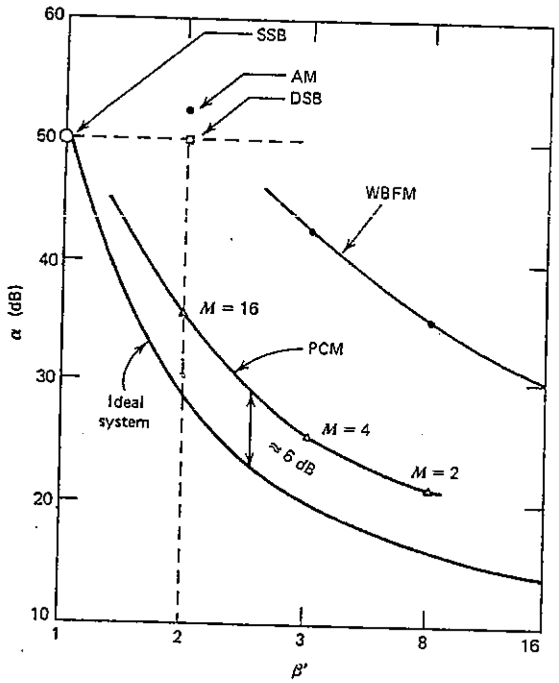


Figure 10.35 Bandwidth and power required for $(S/N)_d$ = 50 dB. $\alpha = (S/N)_r$.

required to produce $(S/N)_d = 50$ dB as a function of bandwidth ratio for various systems. Formulas given in Table 10.3 were used in computing $(S/N)_r$ for AM, DSB, SSB, FM, and the ideal system. For PCM systems, the minimum $(S/N)_r$ needed to produce $(S/N)_d = 50$ dB is calculated by arbitrarily defining the PCM threshold as the point at which symbol errors due to channel noise occur with a probability $P_e < 10^{-4}$. For an M-ary baseband PCM system, P_e is obtained from Equation (5.56) with $S_T = S_r$ as

$$P_e = 2\left(\frac{M-1}{M}\right)Q\left(\sqrt{\frac{6S_r}{(M^2-1)r_s\eta}}\right)$$

where S_r is the average signal power at the receiver input, $r_s = 2f_x \log_M(q)$ is the channel symbol rate, and q is the number of quantizer levels.* For $P_e < 10^{-4}$, we need

$$Q\left(\sqrt{\frac{3\alpha}{(M^2-1)\log_M q}}\right) < \frac{M}{2(M-1)}(10^{-4})$$

If z_0 satisfies $Q(z_0) = M(10^{-4})/2(M-1)$, then we have

$$\alpha \geqslant \left(\frac{(M^2-1)\log_M(q)}{3}\right)z_0^2 \tag{10.61}$$

Above threshold, $(S/N)_d = q^2$, and for $(S/N)_d = 10^5$ we need $q \approx 316$. Knowing the value of q, we can compute β' as

$$r_s = (2f_x)\log_M(q),$$
$$B_T = r_s/2 = f_x \log_M(q)$$

or

$$B_T/f_x = \beta' = \log_M(q) \tag{10.62}$$

Values of α and β' for $M = 2$, 4, and 16 are shown plotted in Figure 10.35 for the PCM system.

The plots in Figure 10.35 show that the power-bandwidth exchange in PCM is considerably better than wideband FM. The PCM system requires about 6 dB more power than the ideal system. In summary, we can say that FM and PCM offer wideband noise reduction and PCM is somewhat better than FM systems at low input signal-to-noise ratios. The performance of all practical systems from a power-bandwidth viewpoint is an order of magnitude below the performance of the ideal system. At low input signal-to-noise ratios SSB and DSB are better than other practical modulation schemes that suffer from threshold effects.

*q is used to denote the number of quantizer levels rather than Q, since Q is used here to denote the area under a normal pdf.

10.7 SUMMARY

We discussed several schemes for transmitting analog message signals using digital transmission techniques. Methods of sampling, quantizing, and encoding analog message signals were discussed. Pulse code modulation and delta modulation schemes were analyzed and the effects of quantizing noise and thermal noise in these systems were discussed. Finally, the performance of PCM, FM, SSB, DSB, and AM systems were compared with the performance of an ideal system. The results derived in this chapter clearly indicate that PCM can be used effectively for trading bandwidth for power. Also, PCM can be used for time division multiplexing a number of analog message signals.

REFERENCES

Additional discussion of the noise performance of PCM and DM systems may be found in books by Panter (1965) and Cattermole (1969). Recent techniques for digital transmission of voice and data are summarized in two articles published in the IEEE Spectrum (1977) and the IEEE Proceedings (1977).

The minimum mean squared error quantizer design discussed in Section 10.3.2 is based on Max's work (1960). His original paper contains tables of quantizer end points and output levels for several values of Q.

1. J. Max, "Quantizing for Minimum Distortion." *IRE Transactions on Information Theory*, Vol. IT-6 (1960), pp. 7–12.
2. A. Papoulis. *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, New York (1965).
3. "Optical Transmission of Voice and Data," and "Digital Telephones." *IEEE Spectrum*, Vol. 14, February (1977).
4. P. F. Panter. *Modulation, Noise and Spectral Analysis*. McGraw-Hill, New York (1965).
5. K. W. Cattermole. "Principles of Pulse Code Modulation." *American Elsevier*, New York (1969).
6. *The Philosophy of PCM*. Bell System Monograph; (also in *PROC. IRE* (Nov. 1948))
7. B. Gold. "Digital Speech Networks." *IEEE Proceedings*, Vol. 65, December (1977).

PROBLEMS

Section 10.2

10.1. A lowpass signal $x(t)$ has a spectrum $X(f)$ given by

$$X(f) = \begin{cases} 1 - |f|/200, & |f| < 200 \\ 0, & \text{elsewhere} \end{cases}$$

(a) Assume that $x(t)$ is ideally sampled at $f_s = 300$ Hz. Sketch the spectrum of $x_\delta(t)$ for $|f| < 200$.
(b) Repeat part (a) with $f_s = 400$ Hz.

10.2. A signal $x(t) = 2\cos 400\pi t + 6\cos 640\pi t$ is ideally sampled at $f_s = 500$ Hz. If the sampled signal is passed through an ideal lowpass filter with a cutoff frequency of 400 Hz, what frequency components will appear in the output?

10.3. A bandpass signal with a spectrum shown in Figure 10.36 is ideally sampled. Sketch the spectrum of the sampled signal when $f_s = 20$, 30, and 40 Hz. Indicate if and how the signal can be recovered.

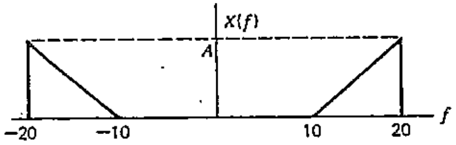


Figure 10.36 Signal spectrum for Problem 10.3.

10.4. A bandpass signal with a spectrum shown in Figure 10.37 is ideally sampled.
(a) Show that the signal can be reconstructed when $f_s = f_{xu} = 2.5B_x$.
(b) Show that the signal can be reconstructed when $f_s > 2f_{xu} = 5B_x$.
(c) Show that aliasing takes place when $f_s = 3.5B_x$.

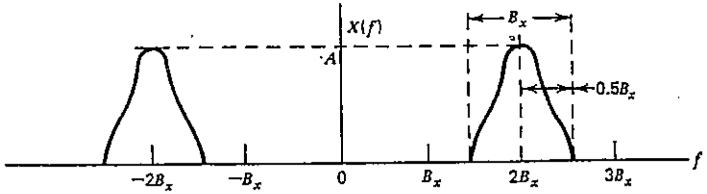


Figure 10.37 Signal spectrum for Problem 10.4.

10.5. Consider the signal $x(t) = e^{-\alpha t}u(t)$, which is not bandlimited. Determine the minimum sampling rate (in terms of the number of −3 dB bandwidths of $x(t)$) such that the magnitude of the largest aliasing frequency component introduced by sampling is at least 10 dB below the magnitude of the largest spectral component of $x(t)$.

10.6. A rectangular pulse waveform is sampled once every T_s seconds and reconstructed using an ideal LPF with a cutoff frequency of $f_s/2$ (see Figure 10.38). Sketch the reconstructed waveform for $T_s = \frac{1}{6}$ sec and $T_s = \frac{1}{12}$ sec.

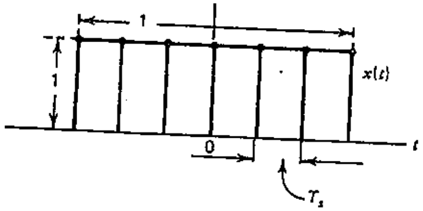


Figure 10.38 Sampling of a rectangular waveform.

10.7. The uniform sampling theorem says that a bandlimited signal $x(t)$ can be completely specified by its sampled values in the time domain. Now, consider a time limited signal $x(t)$ that is zero for $|t| \geqslant T$. Show that the spectrum $X(f)$ of $x(t)$ can be completely specified by the sampled values $X(kf_0)$, where $f_0 \leqslant 1/2T$.

10.8. Show that $\Sigma_{k=-\infty}^{\infty} x(kT_s) = f_s \Sigma_{m=-\infty}^{\infty} X(mf_s)$, where $x(t)$ is bandlimited to f_x and $f_s = 2f_x$.

Section 10.3

10.9. The probability density function of the sampled values of an analog signal is shown in Figure 10.39. Design a four-level uniform quantizer and calculate the signal to quantizing noise power ratio.

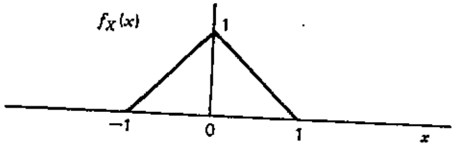


Figure 10.39 Signal pdf for Problems 10.9, 10.10, and 10.11.

10.10. (a) For the pdf shown in Figure 10.39, design a four-level minimum mean squared error nonuniform quantizer.
(b) Compute the signal to quantizing noise power ratio for the nonuniform quantizer.
(c) Design a compressor and expander so that the nonuniform quantizing can be done using a compressor and uniform quantizer.

10.11. Redraw the compressor shown in Figure 10.15, using a piecewise linear approximation. Let $y = g(x)$ be the resulting transfer characteristic and let $m_1, m_2, \ldots, m_6$ be the midpoints of the intervals on the x axis, and let the $\Delta_1, \Delta_2, \ldots, \Delta_6$ be the step sizes.

(a) Show that the piecewise linearity assumption implies

$$\Delta_i = \Delta / g'(m_i)$$

where Δ is the output step size and

$$g'(m_i) = \left.\frac{dg(x)}{dx}\right|_{x=m_i}$$

(b) If $f_X(x)$ is approximately constant throughout the step, show that the quantizing noise power is given by

$$N_q \approx \frac{1}{12} \sum_{i=1}^{6} (\Delta_i)^3 f_X(m_i)$$

(c) Using the result of (b), show that if the number of quantization levels is large, then

$$N_q \approx \frac{\Delta^2}{12} \sum_i \frac{\Delta_i f_X(m_i)}{[g'(m_i)]^2} \approx \frac{\Delta^2}{12} \int_{x_{\min}}^{x_{\max}} \frac{f_X(x)\,dx}{[g'(x)]^2}$$

(d) The *companding improvement* factor c_f is defined as the ratio of the quantizing noise power with no companding to the quantization error with companding. Obtain an expression for c_f.

10.12. The logarithmic compression used in processing speech signals has the characteristic

$$y = \begin{cases} x_{\max} \dfrac{\log_e(1 + \mu x/x_{\max})}{\log_e(1+\mu)}, & 0 \leqslant x \leqslant x_{\max} \\ -x_{\max} \dfrac{\log_e(1 - \mu x/x_{\max})}{\log_e(1+\mu)}, & -x_{\max} \leqslant x \leqslant 0 \end{cases}$$

(a) Sketch the compression characteristic with $\mu = 0, 5, 10, 100$.

(b) Plot the corresponding expander characteristics.

(c) If there are 64 quantization levels, discuss the variation of step size versus the input voltage x.

(d) Assuming X to be uniformly distributed between $-x_{\max}$ to $x_{\max}$, show that the companding improvement factor is

$$c_f = \left(\frac{\mu}{\log_e(1+\mu)}\right)^2 \left(\frac{1}{1 + \mu + \mu^2/3}\right)$$

10.13. Signals are sometimes quantized using an equal probability quantizing (maximum entropy) algorithm wherein the quantizer levels are made to occur with equal probability, that is, $P(X_q = m_i) = 1/Q$ for $i = 1, 2, 3, \ldots, Q$.

(a) Design a four-level equal probability quantizer for the pdf shown in Figure 10.39.

(b) Compare the signal to quantizing noise power ratio of the equal probability quantizer with that of the minimum mean squared error quantizer (Problem 10).

(c) Compare the entropies of the output levels for the equal probability quantizer and the minimum mean squared error quantizer.

10.14. Nonbandlimited signals are usually filtered before being sampled (see Figure 10.40a). Filtering distorts the signal even without quantizing (see Figure 10.40b).

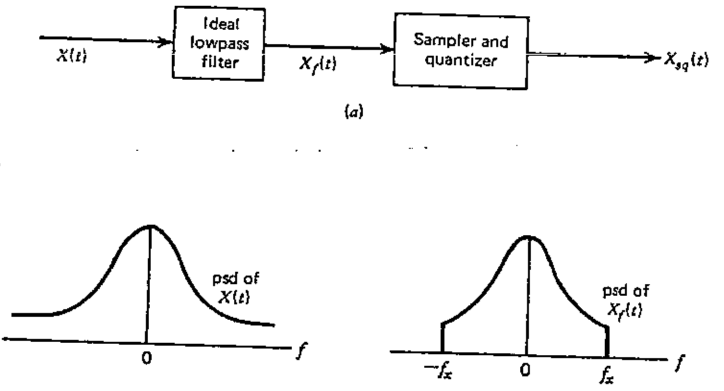
(a) Show that the distortion due to filtering, N_d, is given by

$$N_d = E\{[X(t) - X_f(t)]^2\}$$
$$= 2\int_{f_x}^{\infty} G_X(f)\,df$$

(b) Assume that $X(t)$ has a Gaussian pdf and

$$G_X(f) = e^{-|f/f_1|}$$

If $X(t)$ is sampled at a rate $f_s = 2f_x$, find the output signal-to-noise



(a)

(b)

Figure 10.40 (a) Quantizing of nonbandlimited signals. (b) Effect of filtering.

ratio

$$\left(\frac{S}{N}\right)_d = \frac{S_0}{N_d + N_q}$$

where N_q is the average noise power due to quantizing and S_0 is the average signal power at the quantizer output. [*Hint:* For random signals,

$$\begin{aligned} E\{[X(t) - X_f(t)]^2\} &= R_{XX}(0) + R_{X_fX_f}(0) - 2R_{XX_f}(0) \\ &= \int_{-\infty}^{\infty} G_X(f)\, df + \int_{-\infty}^{\infty} G_X(f)|H(f)|^2\, df \\ &\quad - 2\int_{-\infty}^{\infty} G_X(f)H(f)\, df \end{aligned}$$

where $H(f)$ is the filter transfer function.]

10.15. Consider the differential quantizer shown in Figure 10.41. The signal to be quantized is a zero mean Gaussian random process with an autocorrelation function

$$R_{XX}(\tau) = e^{-6000|\tau|}$$

The signal is sampled at a rate of 12,000 Hz and differentially quantized using a minimum mean squared error quantizer. The error due to quantizing can be approximated by

$$E\{[Y(kT_s) - Y_q(kT_s)]^2\} = (2.2)\sigma_Y^2 Q^{-1.96} = N_q$$

and the performance of the differential quantizer is measured by the signal to quantizing noise power ratio defined as

$$\frac{S_0}{N_q} \triangleq \frac{\sigma_X^2}{N_q}$$

where Q is the number of quantizer levels.

(a) With $\hat{X}(kT_s) = X[(k-1)T_s]$, find the minimum number of quantizer levels needed to obtain a signal to quantizing noise power ratio of 40 dB.

(b) Repeat (a) with

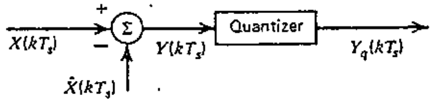


Figure 10.41 Differential quantizer.

$$\hat{X}(kT_s) = \frac{R_{XX}(T_s)}{R_{XX}(0)} X[(k-1)T_s]$$

(c) Repeat with $\hat{X}(kT_s) = 0$ (i.e., direct quantizing).

10.16. Repeat 10.15 (a) and (b) with $f_s = 24{,}000$ Hz.

Section 10.4

10.17. The threshold value of the input signal-to-noise ratio $(S/N)_i$ in PCM systems is defined as the value of $(S/N)_i$ for which the value of $(S/N)_0$ is 1 dB below its maximum.

(a) Show that the threshold occurs when

$$P_e \approx 1/[(16)2^{2N}]$$

(b) Plot P_e versus N, for $N = 2, 4, 6, 8$, and 10.

(c) Assuming that a PSK signaling scheme is used, sketch the threshold values of $(S/N)_i$ versus N for $N = 2, 4, 6, 8$, and 10.

10.18. A signal $X(t)$ bandlimited to 15 kHz is sampled at 50 kHz and the samples are transmitted using PCM/PSK. An output S/N of at least 40 dB is desired. (Assume that $X(t)$ has a uniform pdf).

(a) Find the bandwidth requirements of the system.

(b) Find $(S/N)_i$ if the system is to operate above threshold.

(c) Find $(S/N)_0$ if the system is operating with a $(S/N)_i$ that is 3 dB below the threshold value.

10.19. A nonbandlimited signal $X(t)$ has a power spectral density

$$G_X(f) = e^{-|f/3000|}$$

The signal is bandlimited to f_x Hz by ideal lowpass filtering.

(a) Find the value f_x such that the filter passes at least 90% of the signal power at its input.

(b) If the filter output is converted to 6-bit PCM and transmitted over a binary symmetric channel with $P_e = 10^{-5}$, find the overall signal-to-noise power ratio defined as

$$\left(\frac{S}{N}\right)_0 = \frac{S_0}{N_d + N_q + N_{th}}$$

N_d, N_q, and N_{th} are the distortion due to filtering and the average noise power due to quantizing, and channel bit errors, respectively.

10.20. The output of an analog message source is modeled as a zero mean random process $X(t)$ with a uniform pdf and the power spectral

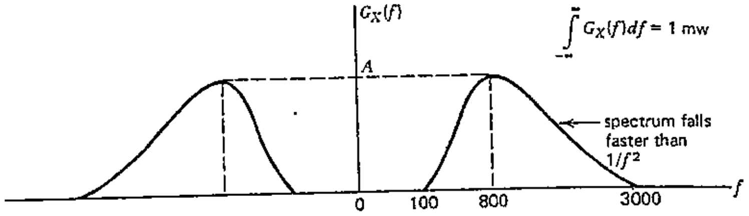


Figure 10.42 Psd of the output of an analog information source.

density shown in Figure 10.42. This signal is to be transmitted using DM.

(a) Find the sampling rate and the step size required to maintain a signal to quantizing noise power ratio of 40 dB.

(b) Compare the bandwidth of the DM with a PCM system operating with the same signal to quantizing noise power ratio. Assume that $f_s = 6000$ Hz for the PCM system.

10.21. The threshold value of P_e in DM systems is defined as the value of P_e for which the value of $(S/N)_0$ is 1 dB below its maximum. Express the value of P_e in terms of f'_s, f_x, and f_1.

10.22. Compare the threshold values of P_e for the DM and PCM systems discussed in Problem 10.20. Assuming a PSK signaling scheme, find the threshold value of the signal-to-noise ratios at the input for the DM and PCM receivers.

10.23. Plot $(S_0/N_q)_{\mathrm{DM}}$ and $(S_0/N_q)_{\mathrm{PCM}}$ versus B_T for equal transmission bandwidths $B_T = 4f_x$ to $32f_x$. Assume $f_x = 3000$ Hz and $f_1 = 300$ Hz.

Section 10.5

10.24. Two lowpass signals of equal bandwidth are impulse sampled and time multiplexed using PAM. The TDM signal is passed through a lowpass filter and then transmitted over a channel with a bandwidth of 10 kHz.

(a) What is the maximum sampling rate for each channel to insure that each signal can be recovered at the receiver?

(b) What is the maximum frequency content allowable for each signal?

(c) Sketch a block diagram for the transmitter and receiver.

10.25. Eight input signals are sampled and time multiplexed using PAM. The time multiplexed signal is passed through a lowpass filter before transmission. Six of the input signals have a bandwidth of 4 kHz and the other two are bandlimited to 12 kHz.

(a) What is the minimum overall sampling rate if all channels are sampled at the *same* rate?

(b) Design an asynchronous TDM for this application.

(c) Compare the transmission bandwidth requirements of (a) and (b).

10.26. Twenty-four analog signals, each having a bandwidth of 15 kHz, are to be time-division multiplexed and transmitted via PAM/AM. A guard band of 5 kHz is required for signal reconstruction from the PAM samples of each signal.

(a) Determine the sampling rate for each channel.

(b) Determine the transmission bandwidth.

(c) Draw functional block diagrams of the transmitter and receiver.

10.27. A number of 20 kHz channels are to be time-division multiplexed and transmitted using PAM. The sampling pulses are 4 μsec in width. The TDM pulse train is passed through a lowpass filter with a time constant $RC = 1$ μsec. (In previous examples we assumed that this baseband filtering was done by an ideal lowpass filter.) This filtering introduces crosstalk (or intersymbol interference) between channels, that is, a certain amount of signal energy from one pulse "spills" over into the time slot of the next pulse. Define the crosstalk factor as the ratio of signal energy from a pulse that spills over into the next time slot and the signal energy within the time slot allotted for the pulse. Using this criteria:

(a) Find the crosstalk factor for five channels.

(b) If the crosstalk factor is to be less than 0.01, find the pulse width for a five-channel system.

Section 10.6

10.28. Is it possible to design a communication system to yield $(S/N)_d = 60$ dB, with $(S/N)_r = 20$ dB, and $\beta' = 4$?

10.29. What is the minimum bandwidth expansion ratio required to obtain $(S/N)_d = 60$ dB with $(S/N)_r = 20$ dB?

10.30. A video signal having a bandwidth of 6 MHz is to be transmitted from the moon using 8-bit PCM/PSK. The thermal noise power spectral density at the receiving antenna is $\eta/2 = 10^{-12}$ watt/Hz. Assuming a power loss of 40 dB in the channel, calculate the threshold power requirements of the transmitter on the moon and compare it with the power requirements of an ideal system using the same bandwidth as the PCM system.

10.31. Suppose that black and white still pictures from the moon were digitized using $(25)(10^4)$ samples per picture (500×500 array) and transmitted to the earth using 8-bit PCM/PSK. Assume that 10 minutes are allotted for