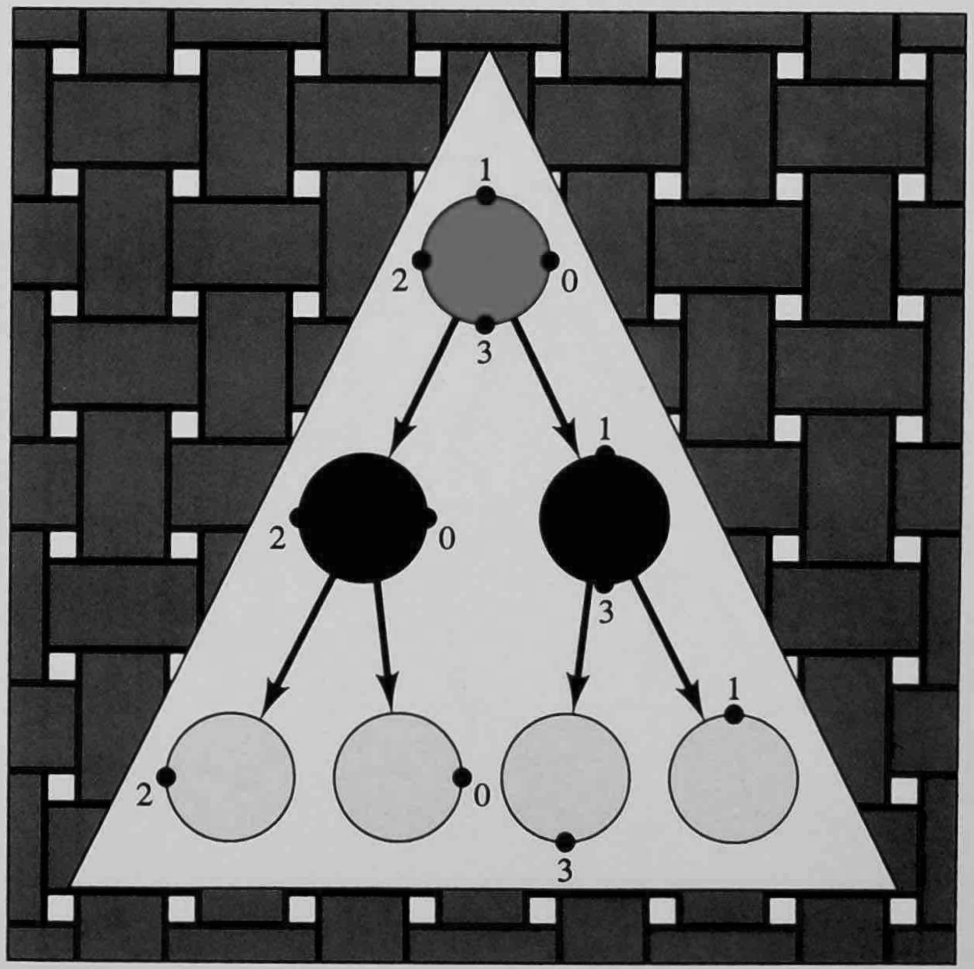
Introduction to Trellis-Coded Modulation with Applications

Ezio Biglieri
Dariush Divsalar
Peter J. McLane
Marvin K. Simon



Introduction to Trellis-Coded Modulation with Applications

Ezio Biglieri
Politecnico di Torino, Italy

Dariush Divsalar
Jet Propulsion Laboratory
California Institute of Technology

Peter J. McLane
Queen's University, Canada

Marvin K. Simon
Jet Propulsion Laboratory
California Institute of Technology

Macmillan Publishing Company
New York
Maxwell Macmillan Canada, Inc.
Toronto
Maxwell Macmillan International
New York Oxford Singapore Sydney

Editor: John Griffin
Production Supervisor: Elaine W. Wetterau
Production Manager: Pamela Kennedy Oborski
Text Designer: Eileen Burke
Cover Designer: Doris Chen
Cover Illustration: Marvin Simon
Illustrations: Publication Services

This book was set in 10/12 Times Roman by Publication Services, printed and bound by Quinn Woodbine, Inc. The jacket was printed by Phoenix Color Corporation.

Printed in the United States of America

Macmillan Publishing Company
866 Third Avenue, New York, New York 10022

Maxwell Macmillan Canada, Inc.
1200 Eglinton Avenue, E.
Suite 200
Don Mills, Ontario M3C 3N1

Library of Congress Cataloging in Publication Data

Introduction to trellis-coded modulation, with applications/Ezio Biglieri . . . [et al.].
p. cm.
Includes bibliographical references and index.
ISBN 0-02-309965-8
1. Digital modulation. 2. Modulation (Electronics) 3. Coding theory. I. Biglieri, Ezio.
TK5106.7.I58 1991
621.381'536—dc20 91–10991
CIP

Printing: 1 2 3 4 5 6 7 8 9 Year: 1 2 3 4 5 6 7 8 9 0

Contents

CHAPTER 1

Introduction

We begin by outlining the structure of digital communication systems and end the first chapter with an outline of Shannon's information theory, which includes some aspects of elementary channel and source coding. The second chapter contains the rest of our material on traditional coding theory: a consideration of BCH codes and some material on convolutional codes. The main intent of the first two chapters is to supply the reader with the necessary background in information theory and error correction coding to understand the theory of trellis-coded modulation. Another goal is to provide the essentials of information theory and coding where the book is used for a single course on these subjects that contains a significant component on trellis-coded modulation.

1.1 Digital Communications Structure

Digital communications systems have a definite structure and knowledge of this structure is helpful in understanding the role of coding and modulation systems. The simplest structure, shown in Fig. 1.1, is for a point-to-point communication system—not that for a communications network or a point-to-multipoint system. We have a transmitter, T_x, a receiver, R_x, and a channel that links the transmitter and receiver.

The transmitter, channel, and receiver shown in Fig. 1.1 can be further subdivided. Let us begin by considering a subdivision of the transmitter structure (Fig. 1.2). We have an information source that we will take as binary, which means that its output is a sequence in which the only elements are 1s and 0s. The source is followed by a source encoder, a channel encoder, and a modulator. We now describe the function of each of these entities. Note that if the source is analog—for example, a speech or video source—we shall assume that it has been digitized.

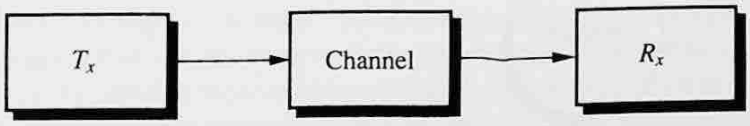


FIGURE 1.1 Simplest digital communications structure.

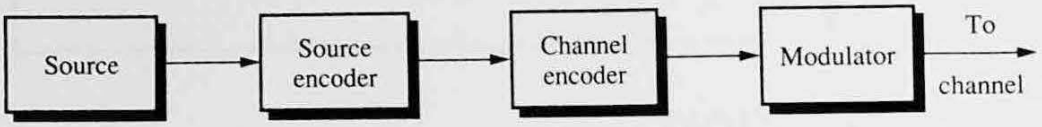


FIGURE 1.2 Transmitter block diagram.

1.1.1 Source Encoder

A good, or desirable source is random—such sources have maximum information. Clearly, if a 1 and a 0 have an equal probability of occurrence, knowledge of the source output provides a maximum amount of information. For instance, if the source nearly always outputs a 1, its output can be predicted and knowledge of the output provides very little information. Usually, sources are not random and contain significant amounts of redundancy. For example, in a video image, neighboring picture elements are usually strongly related. The role of a source encoder is to randomize the source. A measure of randomness is entropy, a concept borrowed from thermodynamics. The function of a source encoder, then, is as illustrated in Fig. 1.3.

Why do we want the source to be encoded to a disordered state? The answer lies in the utilization of one of the scarce resources of the telecommunications problem—the channel. We should not waste the scarce resources of the channel by sending predictable quantities over this link between the receiver and the transmitter. The channel should only be used to carry the unpredictable information from the source, that is, the output from the source encoder.

1.1.2 Channel Encoder

The goal of the channel encoder is to introduce an error correction capability into the source encoder output to combat channel transmission errors. To achieve this goal, some redundancy must be added to the source encoder output. This may seem confusing at first because we have just argued that all redundancy must be stripped from the source outputs for efficient channel transmission. Indeed, this book is about a technique, trellis coding, for adding redundancy to the source outputs so that the channel is utilized from a very efficient point of view. However, the first two chapters are about the traditional way of adding redundancy, through parity checks, and then transmitting the information plus parity bits across the channel in

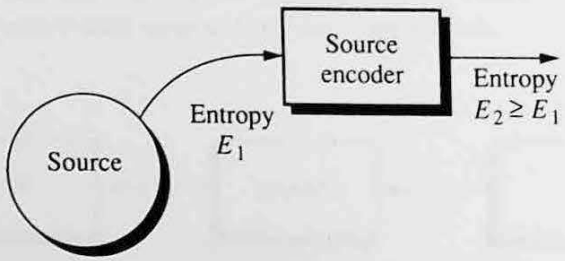


FIGURE 1.3 Function of a source encoder.

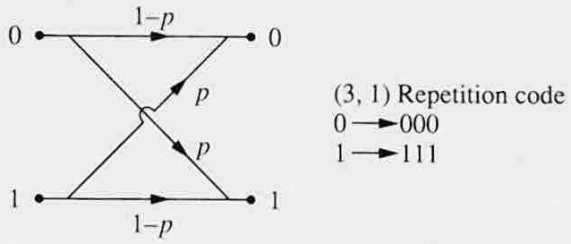


FIGURE 1.4 BSC plus (3, 1) repetition code.

a time-serial manner. Note that the same parity bits will be appended to a unique collection of source output bits called the message. In this way, the redundancy we add to the message is controlled and the receiver will have knowledge of the structure of this redundancy. This is the difference between the original redundancy in the source symbols, which is not controlled, and the redundancy added in channel coding, which is controlled. Let us consider a simple example of channel coding. The binary symmetric channel (BSC) plus a (3, 1) repetition code are illustrated in Fig. 1.4.

The transmission diagram is a summarizing diagram for transmission over the channel, illustrating the fact that transmission errors occur with a prescribed probability of p. The channel coder appends two identical parity bits to the source symbol, and the resulting 3-bit word is transmitted over the channel one bit at a time (in Section 1.1.3 we show how this could be done). We can regard channel transmission as three uses of the BSC shown in Fig. 1.4. If (0, 1, 0) is the 3-bit output from three uses of the BSC, we should declare (0, 0, 0) as transmitted (denoted as t_x), since if 0 was the source bit sent, we have corrected a channel transmission error in position 2. Thus our decoder for the BSC's 3-bit outputs is based on majority rule and so will always result in one channel error being corrected no matter where it appears in the 3-bit word.

The situation described above can be represented using the cube shown in Fig. 1.5. Two possible code words transmitted over the BSC are separated by a Hamming distance, d^{H}, of 3 and nearest-neighbor decoding results on a single error being corrected. In general, if two code words differ in their component position, we add one to their component distance, and examining all components gives the Hamming distance between the two code words. Here we have $d^{\mathrm{H}} = 3$. In general, for more than two code words the greatest chance for error comes in comparing two code words of least distance, $d^{\mathrm{H}}_{\min}$. In addition, the number of errors that can be corrected by a code with the shortest Hamming distance, $d^{\mathrm{H}}_{\min}$, is $t = \lfloor (d^{\mathrm{H}}_{\min} - 1)/2 \rfloor$, where $\lfloor x \rfloor$ is the largest integer less than or equal to x. In the present example we have $t = 1$ correctable channel transmission errors per code word received as $d^{\mathrm{H}}_{\min} = 3$.

The detection of errors is also a key item in channel coding because we could always request, through a feedback channel, retransmission of a code word detected to contain errors. In the present example only one error can be detected: for instance, (0, 1, 1) could be received (denoted as r_x) when (0, 0, 0) was transmitted, but this error pattern is not detectable since the decoder must also consider (1, 1, 1) as a candidate transmitted code word. Clearly, a single channel error can always

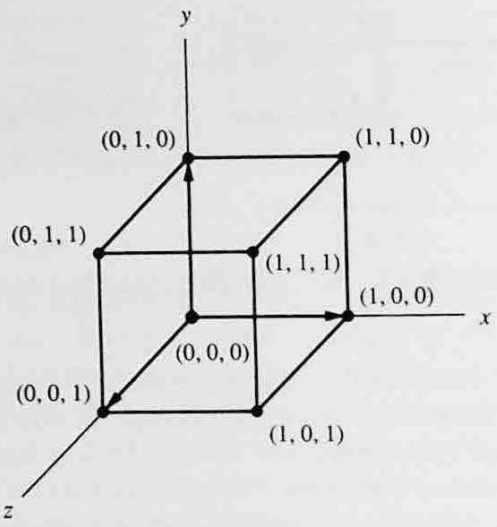


FIGURE 1.5 Decoding represented as points on a cube.

be detected for the present example. In general, $\lfloor d^{\mathrm{H}}_{\min}/2 \rfloor$ transmission errors can always be detected and for the present example we have only one error detected. However, if we used the code $1 \rightarrow (1, 1, 1, 1)$ and $0 \rightarrow (0, 0, 0, 0)$, we would have $d^{\mathrm{H}}_{\min} = 4$, and thus two errors detected but still only one error corrected. Note here that only 1 out of 4 bits sent is an information bit, and we say that the rate of the code is 1/4. The earlier code had rate 1/3 and thus less error detection capability than given in the rate 1/4 case. Our concept of error detection here is different than in most textbooks on coding theory, where the number of errors detected is taken as $d^{\mathrm{H}}_{\min} - 1$. In traditional coding the rate 1/3 code is transmitted by using the BSC three times for each information bit. To realize this in practice, we must either speed up the rate of symbol transmission by a factor of 3 or keep the same rate of symbol transmission and be content with one-third the information transmission rate relative to when no channel coding is used. Thus, in either case, an increased channel bandwidth is required per information bit transmitted. This book is about an alternative to this approach that involves no change in information transmission rate; rather, the number of points in the signal constellation for modulation is increased to achieve the required redundancy.

1.1.3 Modulator

In Fig. 1.2 the modulator interfaces the channel encoder to the channel. The source, source encoder, and channel encoder taken together can be viewed as a modified binary source that feeds the modulator, and as such, the modulator can be regarded as interfacing the source to the channel. Physical channels can require electrical signals, radio signals, or optical signals. The modulator takes in the source outputs and outputs waveforms that suit the physical nature of the channel and are also chosen to yield either system simplicity or optimal detection performance. A baseband binary modulator is shown in Fig. 1.6. We call this a baseband modulator because no sinusoidal carrier signal is involved. On a channel that has symmetric

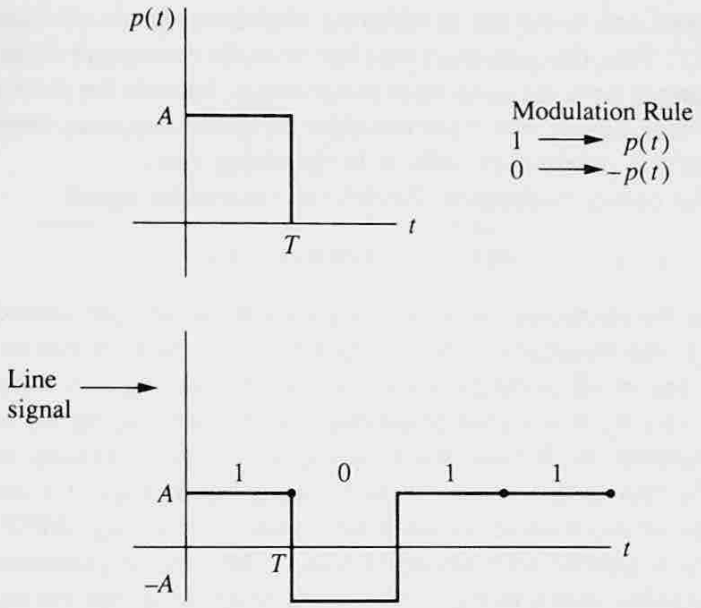
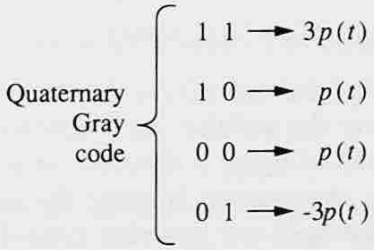


FIGURE 1.6 Baseband modulator.

interference, the signal selection in Fig. 1.6 is optimal in that it will yield for a fixed transmitted power the least number of errors in detection in the receiver. In the quaternary modulator shown in Fig. 1.7, two source bits per symbol interval T are required, whereas before, a single bit will do. In the quaternary case the symbol transmission rate is $2/T$ bits per second (bits/s). This is an example of pulse amplitude modulation; the amount of channel bandwidth such signals



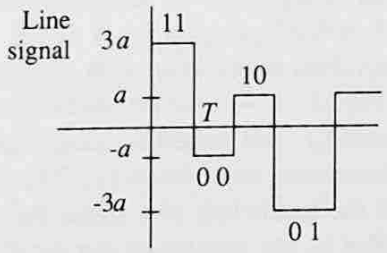


FIGURE 1.7 Quaternary modulator.

require is related only to the rate at which the modulator signals are changed, that is, to the rate $1/T$. Thus the quaternary case has twice the throughput of the binary case and clearly cannot have the same error performance, because the receiver must sort out which of four signals were transmitted for the quaternary case, whereas a signal selection over two possibilities suffices in the binary case.

To consider carrier modulation, consider the sinusoidal signal

$$s(t) = A(t)\cos[\omega_c t + \theta(t)] \tag{1.1}$$

where $A(t)$ is the amplitude; ω_c is the frequency in radians per second and equals $2\pi f_c$, with f_c the frequency in hertz; and $\theta(t)$ is the phase. In carrier modulation we can vary any or all of the parameters (A, ω_c, θ): varying A is called amplitude modulation, varying θ is called phase modulation, and varying ω_c is called frequency modulation; in all cases the variation is (hopefully) linearly related to the message to be transmitted. Some examples are given in Figs. 1.8 and 1.9. Note that binary phase modulation (called binary phase shift keying, BPSK) is equivalent to binary amplitude shift keying (BASK). The type of quadrature amplitude modulation (QAM) shown in Fig. 1.9 is called 64-QAM in that the signal constellation contains 64 points. Thus $6/T$ bits per symbol interval T can be transmitted over the channel. Inherent in the use of this modulation is the fact that two carrier signals that differ in phase by $\pm 90°$ can be separated in the receiver to recover the signals $X(t)$ and $Y(t)$, known as the in-phase and quadrature signals, respectively. This can be done in a coherent receiver, which is a receiver that must acquire and track any nonmodulation phases that exist in the received signal.

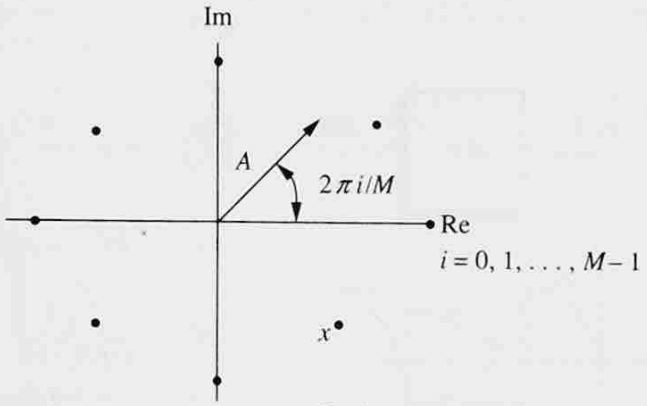
1.1.4 The Communications Channel

The simplest channel is the additive noise channel: here the signal is received with no distortion except additive noise. That is, if $r(t)$ is the received signal,

$$r(t) = s(t) + n(t) \tag{1.2}$$

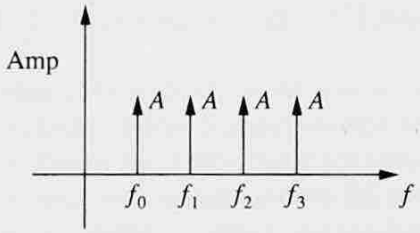
where $s(t)$ is the transmitted signal and $n(t)$ is the additive noise. The classical theory of communication over the additive noise channel is given in reference [1]. A channel where the received signal is distorted, or at least can be distorted, is shown in Fig. 1.10. This phenomenon is called the intersymbol interference channel, as modulation symbols spill over into other symbol intervals, thus causing distortion. Additive noise is also present in the received signal but is not shown on the waveforms in Fig. 1.10.

Define the distribution of signal power as a function of frequency as the power spectrum of a signal. A power spectrum for a QPSK signal is displayed in Fig. 1.11. A QPSK signal involves modulation with a discontinuous phase angle. The other signal spectra in Fig. 1.11—minimum shift keying (MSK), duobinary minimum shift keying (DuMSK), and tamed frequency modulation (TFM)—involve phase modulation with increasing smoothness [2]. This smoothness produces a more compact spectrum. Call the bandwidth of a signal the set of frequencies that contain 98% of its power; that is, the area under the curve over this set of frequencies in Fig. 1.10 that contains 98% of the total area. If the 3-dB bandwidth of the linear



$$s_i(t) = A\cos\left(\omega_i t + \theta + \frac{2\pi i}{M}\right)$$
$$i = 0, 1, \ldots, M-1$$
$$\theta = \text{offset phase}$$

(a)



$$s_i(t) = A\cos(2f_c t + \theta + 2\pi i\Delta t)$$
$$f_i = f_c + i\Delta \qquad i = 0, 1, \ldots, M-1$$

(b)

FIGURE 1.8 Digital phase and frequency modulation: (a) MPSK; (b) FSK.

$$s(t) = X(t)\cos(\omega_c t + \theta) + Y(t)\sin(\omega_c t + \theta)$$

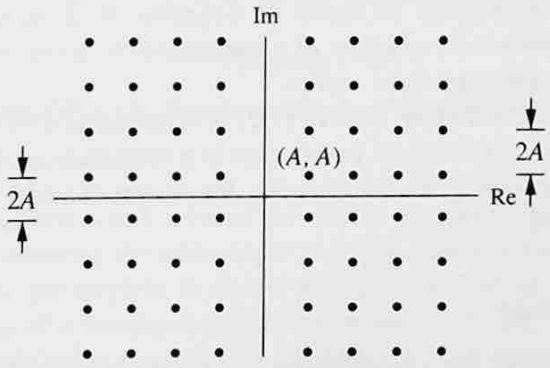


FIGURE 1.9 Quadrature amplitude modulation.

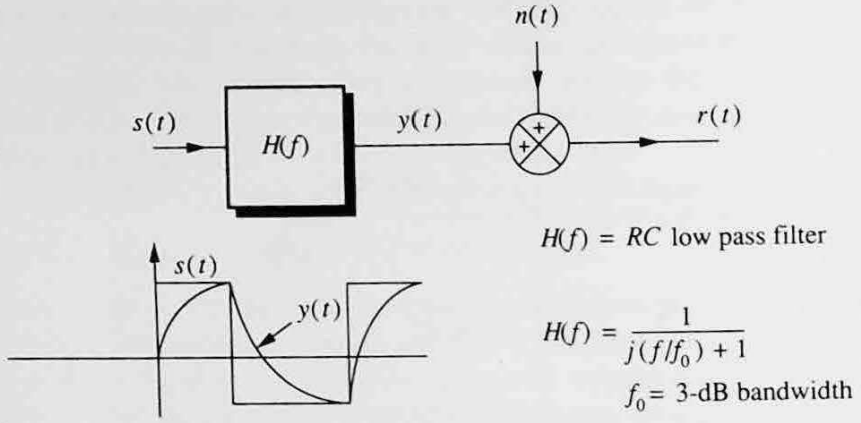


FIGURE 1.10 Intersymbol interference channel.

filter in Fig. 1.9 is significantly less than this bandwidth, intersymbol interference (ISI) results. The classical theory of communication over the ISI channel is given in [3]; a recent textbook [2] considers additive noise, ISI, and some nonlinear channels.

Much of this book is written for modulation and channel coding for the additive noise channel where the additive noise is white Gaussian noise—that is, the signal power spectrum is flat over the bandwidth of all signals sent over the channel. Very little is considered for the ISI channel because the application of trellis codes to this channel is in the early stages of research. A channel that will receive some attention is the nonfrequency selective fading channel (Fig. 1.12). Indeed, the greatest gains in performance that trellis codes have attained are for this channel. Let the input signal be $s(t)$ in equation (1.1); then the output or faded signal is

$$y(t) = G(t)A(t)\cos[\omega_c t + \theta(t) + \psi(t)]. \tag{1.3}$$

In (1.3) the shape of $s(t)$ is not changed; only its amplitude and phase are altered. A typical fading function, $G(t)$, for the model developed in [4] for the Canadian Mobile Satellite Communications System is shown in Fig. 1.13. The classical theory of fading channels in mobile radio systems is given in Jakes's textbook [5], and a good section on fading channels appears in [6]. In addition, material of fading channel models can be found in Appendix A. It should be noted that fading channels represent an example of a multiplicative noise process rather than the additive noise case considered earlier.

In frequency-selective fading, $s(t)$ in (1.2) is distorted as well as attenuated in a time-varying manner. The channel in this case is a combination of the fading channel under consideration and the ISI channel. We do not consider such challenging channels in this book.

1.1.5 The Receiver

The receiver follows the channel in the block diagram in Fig. 1.1. Now the transmitter represents an operation on the source and the function of the transmitter

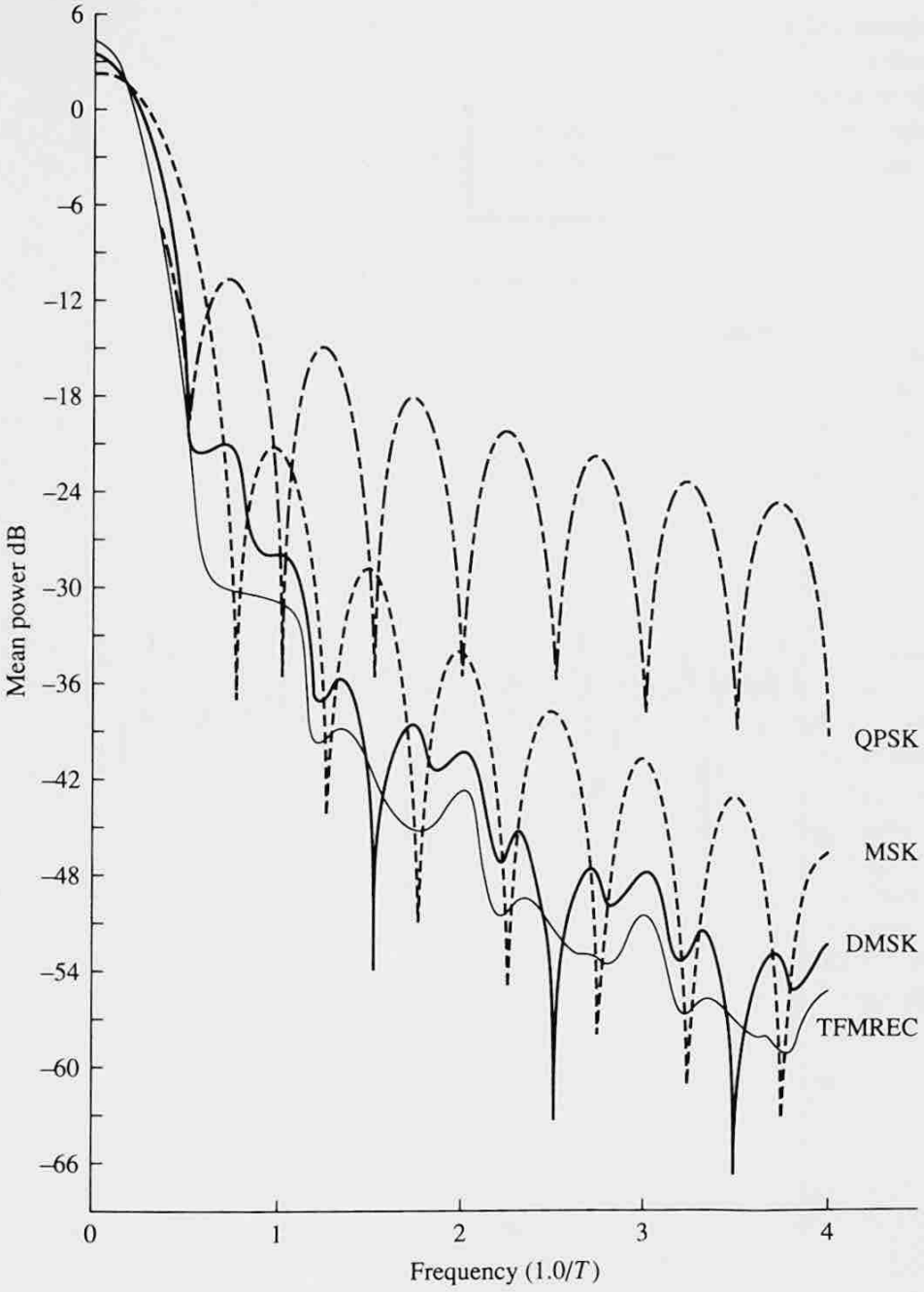
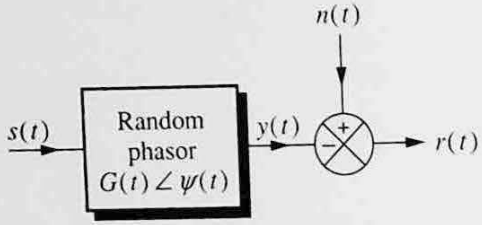


FIGURE 1.11 Power spectrum of various signals.

is to invert this operation and recover the source symbols. Figure 1.2 represents this operation by showing the transmitter in block diagram form. The inverse of this block diagram, the receiver, is shown in Fig. 1.14. Indeed, each block in Fig. 1.14 is the inverse of a corresponding block in Fig. 1.2. The demodulator is the inverse of the modulation process, the channel decoder inverts the channel encoder process, and so on. The various blocks are viewed independently, much as in



$$s(t) = A(t)\cos[\omega_c t + \theta(t)]$$
$$y(t) = G(t)A(t)\cos[\omega_c t + \theta(t) + \psi(t)]$$

FIGURE 1.12 Nonfrequency-selective fading channel.

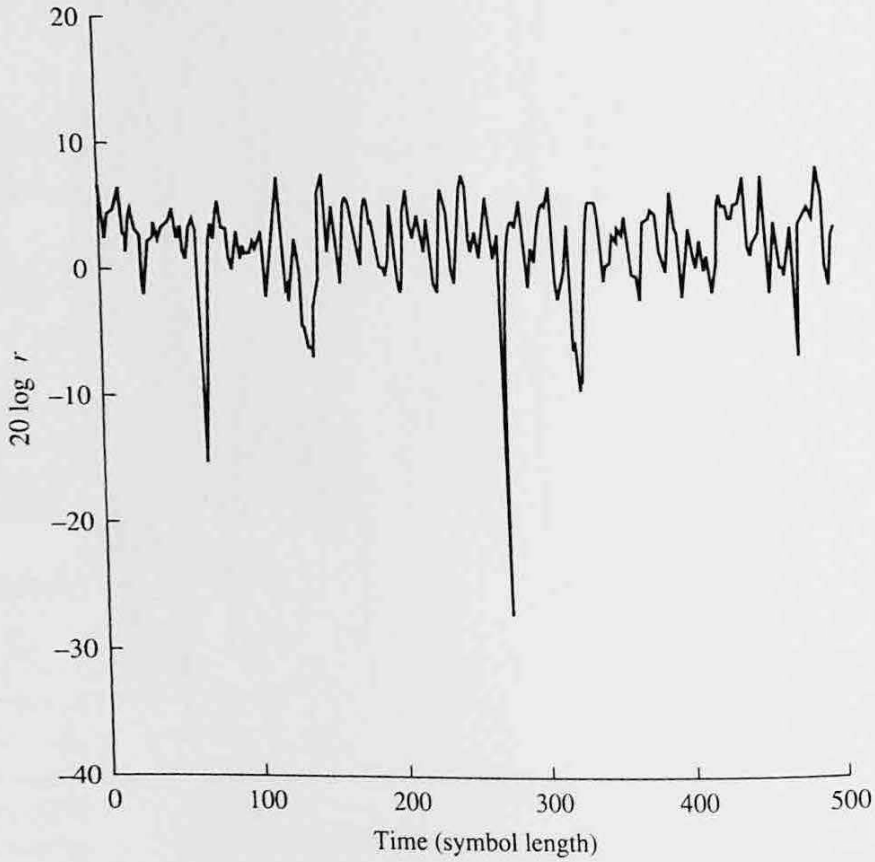


FIGURE 1.13 Amplitude fading function.

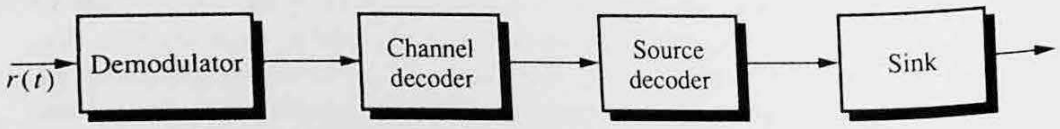


FIGURE 1.14 Receiver block diagram.

the case of multilevel data protocols. Trellis coding serves to merge the processes of modulation and coding, and recent work [7] is aimed at merging the roles of source coding, channel coding, and modulation. We consider an example of a demodulator in the next section. An example of a channel decoder was treated earlier—the majority rule or nearest-neighbor decoder discussed in relation to Fig. 1.4.

1.2 Discrete Memoryless Channels

An example of a discrete memoryless channel (DMC), the BSC, is given in Fig. 1.4. This channel has two inputs and two outputs. In general, a DMC can have a finite number of inputs and outputs. In any case, all of the channels described in Section 1.1.4 involved a continuous-time variable. In this section we show how to derive a DMC from a continuous-time channel description. The latter is a physical channel description, whereas the former is an abstract version of the channel.

1.2.1 Uncoded Baseband Communication

Consider the case of the transmitter, additive noise channel, and receiver shown in Fig. 1.15. The modulation will be as shown in Fig. 1.6, namely, $1 \rightarrow p(t)$ and $0 \rightarrow -p(t)$, where $p(t)$ is the rectangular pulse. The receiver is shown in Fig. 1.16; this is a matched filter and it is optimum in the sense of having the smallest error probability among all receivers [1]. A typical output is shown in Fig. 1.16, together with a sampler that is synchronous with the end of a pulse. These samples are quantized into two levels with a threshold at zero. If the sample is positive, a binary 1 is declared to have been transmitted; otherwise, a binary 0 is declared.

The BSC channel is shown in Fig. 1.4. This channel represents a summary of binary data transmission over the continuous-time channel shown in Fig. 1.15. The BSC is completely described by p, the probability of error per binary digit sent over the channel. Thus, to determine the BSC, we must find p.

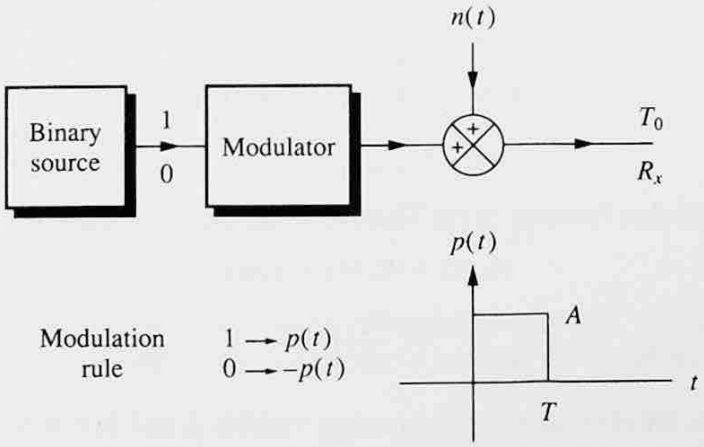


FIGURE 1.15 Baseband system with no coding.

THE FIRST BOOK DEVOTED *COMPLETELY* TO TCM!

Appropriate for students and professionals alike, this book provides a thorough introduction to trellis-coded modulation (TCM), helping readers grasp its theory as well as the techniques needed for its analysis. It offers both a conceptual and practical perspective by applying TCM theory to real-world problems and evaluating the results; examples include fading channels and commercial modems.

Introduction to Trellis-Coded Modulation with Applications contains most of the results of TCM research that have occurred since its invention. In addition, numerical illustrations are included throughout to help describe results from the application of TCM theory.