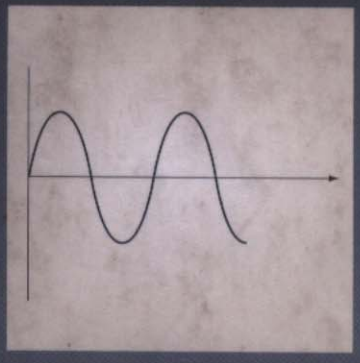
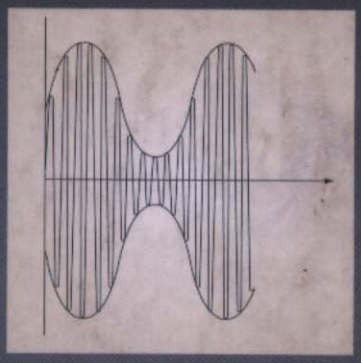
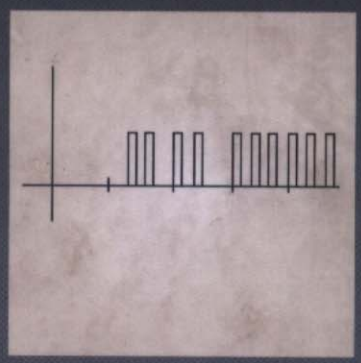
FOURTH EDITION

Analog and Digital Communication Systems







Martin S. Roden

Fourth Edition

Analog and Digital Communication Systems

Martin S. Roden

Department of Electrical and Computer Engineering
California State University, Los Angeles

PRENTICE HALL
Upper Saddle River, New Jersey 07458

Library of Congress Cataloging-in-Publication Data

Roden, Martin S.
Analog and digital communication systems / Martin S. Roden.—4th ed.
p. cm.
Includes bibliographical references and index.
ISBN 0-13-372046-2
1. Telecommunication. 2. Digital communications. I. Title.
TK5105.R64 1996
621.382—dc20 95-11659
CIP

Acquisitions editor: **TOM ROBBINS**
Production editor: **IRWIN ZUCKER**
Copy editor: **BRIAN BAKER**
Cover designer: **BRUCE KENSELAAR**
Buyer: **DONNA SULLIVAN**
Editorial assistant: **PHYLLIS MORGAN**

Printed in the United States of America

10 9 8 7 6

ISBN 0-13-372046-2

Prentice-Hall International (UK) Limited, *London*
Prentice-Hall of Australia Pty. Limited, *Sydney*
Prentice-Hall Canada Inc., *Toronto*
Prentice-Hall Hispanoamericana, S.A., *Mexico*
Prentice-Hall of India Private Limited, *New Delhi*
Prentice-Hall of Japan, Inc., *Tokyo*
Simon & Schuster Asia Pte. Ltd., *Singapore*
Editora Prentice-Hall do Brasil, Ltda., *Rio de Janeiro*

Contents

Preface

The fourth edition of *Analog and Digital Communication Systems* is a greatly modified and enhanced version of the earlier editions. The digital communications portion of the text has been considerably expanded and reorganized. The material is presented in a manner that emphasizes the unifying principles governing all forms of communication, whether analog or digital. Practical design applications and computer exercises have been expanded. In particular, many of the graphs are prepared and formulas are solved using MATLAB™ or Mathcad™. In such cases, the instruction set is presented to give the student practice in using these important tools.

In addition to presenting a unified approach to analog and digital communication, this text strikes a balance between theory and practice. While the undergraduate engineering student needs a firm foundation in the theoretical aspects of the subject, it is also important that he or she be exposed to the real world of *engineering design*. This serves two significant purposes. The first is that an introduction to the real world acts as a strong motivating factor: The "now" generation needs some "touchy-feely" to motivate the spending of hours digging through mathematical formulae. The second purpose of real-world engineering is to ease the transition from academia to the profession. A graduate's first experience with real-world design should not be a great shock, but instead should be a natural transition from the classroom environment.

The book is intended as an introductory text for the study of analog and/or digital communication systems, with or without noise. Although all necessary background material has been included, prerequisite courses in linear systems analysis and in probability are helpful.

The text stresses a mathematical systems approach to all phases of the subject matter. The mathematics used throughout is as elementary as possible, but is carefully chosen so as not to contradict any more sophisticated approach that may eventually be required. An attempt is made to apply intuitive techniques prior to grinding through the mathemat-

ics. The style is informal, and the text has been thoroughly tested in the classroom with excellent success.

Chapter 1, which is for motivational purposes, outlines the environmental factors that must be considered in communicating information. The chapter also clarifies the differences between analog, sampled, and digital signals. The final section lays out a block diagram of a comprehensive communication system. This diagram could serve as a table of contents for the remainder of the text.

Chapter 2 sets forth the mathematical groundwork of signal analysis. It should be review for most students taking the course. Chapter 3 applies signal analysis results to linear systems, with an emphasis on filters. The material in Chapters 2 and 3 is covered in most linear systems courses.

Chapter 4 introduces probability and random analysis and applies these to the study of narrowband noise. The matched filter is introduced as a technique to maximize the signal-to-noise ratio. Chapter 5 deals with baseband communication. Although it concentrates on analog communication, the sampled systems form an important transition into digital communication.

Chapter 6 is a thorough treatment of amplitude modulation (AM), including applications to broadcast radio, television, and AM stereo. Chapter 7 parallels Chapter 6, but for angle instead of amplitude modulation. A section on broadcast frequency modulation (FM) and FM stereo is included. The final section of the chapter compares various angle modulation schemes.

Source encoding is the subject of Chapter 8. In addition to a thorough discussion of the analog-to-digital conversion process, and of the associated round-off errors, the chapter presents baseband forms of digital transmission. Chapter 9 focuses on channel encoding, including data compression, entropy coding, and forward error correction. Both block codes and convolutional codes are analyzed. Baseband forms of digital transmission and reception are the topic of Chapter 10, and modulated forms of transmission are examined in Chapter 11. Transmitters, receivers, error analysis, and timing considerations are treated for amplitude shift keying (ASK), frequency shift keying (FSK), and phase shift keying, (PSK). Hybrid signalling techniques and modems conclude the chapter.

The final chapter, Chapter 12, summarizes important design considerations for both analog and digital communication systems. The chapter concludes with five contemporary case studies.

Problems are presented at the end of Chapters 1 through 11. Numerous solved examples are given within each chapter.

The text is ideally suited for a two-term sequence at the undergraduate level. Chapters 2 and 3 can be omitted if a course in linear systems forms a prerequisite. Chapter 4 can be omitted either if a probability course is a prerequisite or if noise analysis is not part of the course.

If the text is used for a digital communication course, the following sequence of sections and chapters is recommended: Sections 4.7, 5.2, and 5.3, and Chapter 8, Chapter 10, and Chapter 11. Channel encoding (Chapter 9) is normally not included in an introductory digital communication course.

It gives me a great deal of pleasure to acknowledge the assistance and support of many people without whom this text would not have been possible:

To the many classes of students who are responsive during lectures and helped indicate the clearest approach to each topic.

To the faculty around the world who have written to me with comments and suggestions.

To my colleagues at Bell Telephone Labs, Hughes Aircraft Ground Systems, and California State University, Los Angeles. Special thanks go to Professors Roy Barnett, Fred Daneshgaran, George Killinger, and Lili Tabrizi for their many helpful suggestions.

To Dennis J. E. Ross for guidance and assistance.

To Professor A. Papoulis, who played a key role during the formative years of my education.

I sincerely hope this text is the answer to your prayers. If it is, please let me know. If it isn't, please also communicate with me so that, together, we can improve engineering education.

Introduction for the Student

Communication is perhaps the oldest applied area within electrical engineering. As is the case in many technical disciplines, the field of communication is experiencing a revolution several times each decade. Some important milestones of the past include the following:

The television revolution: After only five decades, television has become a way of life. Wristwatch television, wall-size television, and fully interactive cable television are all available.

The space revolution: This has been the catalyst for many innovations in long-distance communication. Satellite communication permits universal access.

The digital revolution: The overriding emphasis on digital electronics and processing has resulted in a rapid change in direction from analog communication to digital communication.

The computer revolution: Microprocessors are changing the shape of everything related to computing and control, including many phases of communication. Home and portable computers (e.g., the "notebook") with modems are popular. They permit direct data communication by the general public.

The consumer revolution: Consumers first discovered calculators and digital watches. They then expanded to sophisticated TV recording systems, CB radio, cellular phones, facsimile (FAX) systems, and video games. When digital watches were not enough, numerous other functions were added to the wrist device, such as calculators, TV, pager receivers, and pulse and temperature indicators.

The personal communication revolution: Cellular telephones began as a tool of the business person. They then expanded into the public sector, with concerns for safety. Instant worldwide analog and digital communication is now desired by a broad spectrum of the public.

There is every reason to expect the rate of revolution to continue or, indeed, accelerate. Two-way interactive cable, high-definition television (HDTV), videotext, and per-

sonal communication will greatly reduce our need for hard copies of literature and for traveling. People are juggling bank accounts, ordering groceries, sending electronic mail, and holding business-related conferences without leaving their homes. The problems of urban centers are being attacked through communication. Indeed, adequate communication can make cities unnecessary. The potential impact of these developments is mind boggling, and the possibilities involving communications are most exciting.

These facts should inspire you to consider the field of communication as a career. But even if you decide not to approach the field any more deeply than in this text, you will be a far more aware person for having experienced even the small amount contained herein. If nothing more, the questions arising in modern communication, as it relates to everything from the space program to home entertainment, will take on a new meaning for you. The devices around you will no longer appear mysterious. You will learn what "magic force" lights the stereo indicator on your receiver, the basics of how a video game works, the inner workings of a FAX machine, and the mechanism by which your voice can travel to any part of the globe in a moment of time.

Enjoy the book! Enjoy the subject! And please, after studying this text, communicate any comments that you may have (positive, negative, or neutral) to me at California State University, Los Angeles, CA 90032. Thank you!

Martin S. Roden
Los Angeles, California

1 Introduction

1.0 PREVIEW

What We Will Cover and Why You Should Care

You will not encounter your first communication system until Chapter 5 of this text. The first four chapters form the framework and define the parameters under which we operate.

Chapter 1 begins with a brief history of the exciting revolution in communication, a revolution that is accelerating at an ever-increasing rate.

We then turn our attention to an investigation of the environment under which our systems must operate. In particular, the characteristics of various communication channels are explored, and we present analytical techniques for the resulting signal distortion. It would not make much sense to start designing communication systems without knowing something about the environment in which they must operate.

The third section of the chapter defines the three types of signals with which we will operate. We begin with analog signals and make a gradual transition to digital signals through the intermediate step of discrete time (sampled) signals.

The block diagram of a communication system is presented and discussed in the final section. The various blocks in the transmitter and receiver form the road map for our excursion through this exciting subject. We present a single comprehensive system block diagram that applies to both analog and digital communication systems. This block diagram forms an abbreviated table of contents for the remainder of the text.

Necessary Background

There are no prerequisites to understanding most of this introductory chapter. The only exception is the discussion of channel distortion in Section 1.2. Understanding the equations in this section requires a basic knowledge of Fourier transforms and systems theory. Such material is reviewed in Chapter 3.

1.1 THE NEED TO COMMUNICATE

Among the earliest forms of communication were vocal-cord sounds generated by animals and human beings, with reception via the ear. When greater distances were required, the sense of sight was used to augment that of hearing. In the second century B.C., Greek telegraphers used torch signals to communicate. Different combinations and positions of torches were used to represent the letters of the Greek alphabet. These early torch signals represent the first example of data communication. Later, drum sounds were used to communicate over greater distances, again calling upon the sense of hearing. Increased distances were possible because the drum sounds were more easily distinguished from background noise than were human vocal-cord sounds.

In the 18th century, communication of letters was accomplished using semaphore flags. Like the torches of ancient Greece, these flags relied on the human eye to receive the signal. This reliance on the eye, of course, severely limited the transmission distances.

In 1753, Charles Morrison, a Scottish surgeon, devised an electrical transmission system using one wire (plus ground) for each letter of the alphabet. A system of pith balls and paper with letters printed on it was used at the receiver.

In 1835, Samuel F. B. Morse began experimenting with telegraphy. Two years later, in 1837, the telegraph was invented by Morse in the United States and by Sir Charles Wheatstone in Great Britain. The first public telegram was sent in 1844, and electrical communication was established as a major component of life. These early forms of communication consist of individual message components such as the letters of the alphabet. (We would later call it digital communication.) It was not until Alexander Graham Bell invented the telephone in 1876 that analog electrical communication became common.

Experimental radio broadcasts began about 1910, with Lee De Forest producing a program from the Metropolitan Opera House in New York City. Five years later, an experimental radio station opened at the University of Wisconsin in Madison. Stations WWJ in Detroit and KDKA in Pittsburgh were among the first to conduct regular broadcasts, in the year 1920.

Public television had its beginning in England in 1927. In the United States, it started three years later. During the early period, broadcasts did not follow any regular schedule. Regular scheduling did not begin until 1939, during the opening of the New York World's Fair.

Satellite communication was launched in the 1960s, with Telstar I being used to relay TV programs starting in 1962 and the first commercial communications satellites being launched in the mid-1960s.

The 1970s saw the beginning of the computer communication revolution. Data transfer is an integral part of our daily lives and has led to a merging of the disciplines of *communication and computer engineering*. *Computer networking* is one of the fastest growing areas of communication.

The *personal communication* revolution began in the 1980s. Before the decade of the nineties is over, the average professional will have a cellular telephone in the car, a portable telephone (no larger than the Star Trek® communicators of the original series), a paging system, a modem in the home computer for use in paying bills or accessing the daily news, and a home FAX machine. Consumers will use fully interactive compact disk

technology, laptops will be networked with worldwide data services, and the global-positioning satellite (GPS) will assist in navigating cars through traffic jams.

The coming millennium is certain to bring a new set of applications and innovations as communication continues to have a significant impact on our lives.

1.2 THE ENVIRONMENT

Before we can begin designing systems to communicate information, we need to know something about the channel through which the signals must be transmitted. We start by exploring the ways in which the channel can change our signals, and then we discuss some common types of channels.

1.2.1 Distortion

Anything that a channel does to a signal other than delaying it and multiplying it by a constant is considered to be *distortion.* (See Chapter 3 for a discussion of *distortionless linear systems.*) Let us assume that the channels we will encounter are *linear* and therefore cannot change the frequencies of their input. Some *nonlinear* forms of distortion are significant at higher transmission frequencies. Indeed, the higher frequencies are affected by air turbulence, which causes a *frequency variation. Doppler radar systems* used for monitoring weather capitalize on this phenomenon.

Linear distortion can cause problems in pulse transmission systems of the type used in pulse modulation or in digital communication. This distortion is characterized by *time dispersion* (spreading), due either to multipath effects or to the characteristics of the channel. For now, we look at the effects that can be readily characterized by the system function of the channel. The channel can be characterized by a system transfer function of the form

$$H(f) = A(f)e^{-j\theta(f)} \tag{1.1}$$

The *amplitude factor* is $A(f)$, and the *phase factor* is $\theta(f)$.

Distortion arises from these two frequency-dependent quantities as follows: If $A(f)$ is not a constant, we have what is known as *amplitude distortion;* if $\theta(f)$ is not linear in f, we have *phase distortion.*

Amplitude Distortion

Let us first assume that $\theta(f)$ is linear with frequency. The transfer function is therefore of the form

$$H(f) = A(f)e^{-j2\pi f t_0} \tag{1.2}$$

where the phase proportionality constant has been denoted as t_0 because it represents the channel delay.

One way to analyze Eq. (1.2) is to expand $A(f)$ into a series—for example, a Fourier series. This can be done if $A(f)$ is bandlimited to a certain range of frequencies. In such cases, we can write

$$H(f) = \sum_{n=0}^{\infty} H_n(f) \tag{1.3}$$

where the terms in the summation are of the form

$$H_n(f) = a_n \cos\left(\frac{n\pi f}{f_m}\right) e^{-j2\pi f t_0} \tag{1.4}$$

These terms are related to the *cosine filter,* whose amplitude characteristic follows a cosine wave in the passband. This filter is shown in Fig. 1.1 for $n = 2$. The system function for this filter is

$$\begin{aligned} H(f) &= \left(A + a\cos\frac{2\pi}{f_m} f\right) e^{-j2\pi f t_0} \\ &= Ae^{-j2\pi f t_0} + \frac{a}{2}\exp\left[j2\pi f\left(\frac{1}{f_m} - t_0\right)\right] \\ &\quad + \frac{a}{2}\exp\left[j2\pi f\left(-\frac{1}{f_m} - t_0\right)\right] \end{aligned} \tag{1.5}$$

Computer Exercise:

Plot Eq. (1.4) for representative values of f_m and t_0.

Solution: We present the instruction steps both for Mathcad and for MATLAB.

Mathcad: We illustrate the instructions to type and the resulting expression on the screen:

```
fm:1                    fm:= 1
f:-2,-1.95;2            f:=-2,-1.95..2
H(f):cos(2*π*f/fm)      H(f):=cos(2.π.f/fm)
```

NOTES: We set fm to unity and step the frequency from -2 to $+2$ in steps of 0.05. Enter π by pressing CONTROL+P. You then enter the plotting mode by pressing "@". Insert "f" for the abscissa and "H(f)" for the ordinate, and the plot results.

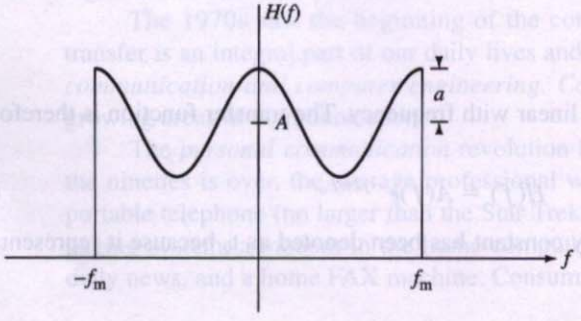


Figure 1.1 Cosine filter.

MATLAB: The following instructions are typed:

```
fm=1
f=-2:.05:2
H=cos(2*pi*f/fm)
plot(f,H)
```

The result is shown in Fig. 1.2.

If the input $r(t)$ to the cosine filter is bandlimited, the output is

$$s(t) = Ar(t - t_0) + \frac{a}{2}r\left(t + \frac{1}{f_m} - t_0\right) + \frac{a}{2}r\left(t - \frac{1}{f_m} - t_0\right) \tag{1.6}$$

Equation (1.6) indicates that the response is in the form of an undistorted version of the input, added to two time-shifted versions *(echoes,* or a *multipath).*

Returning to the case of a general filter, we see that the output of a system with amplitude distortion is a sum of shifted inputs. Thus, with

$$H(f) = \sum_{n=0}^{\infty} a_n \cos\left(\frac{n\pi f}{f_m}\right) e^{-j2\pi f t_0} \tag{1.7}$$

the output due to an input $r(t)$ is

$$s(t) = \sum_{n=0}^{\infty} \frac{a_n}{2}\left[r\left(t + \frac{n}{2f_m} - t_0\right) + r\left(t - \frac{n}{2f_m} - t_0\right)\right] \tag{1.8}$$

Equation (1.8) can be computationally difficult to evaluate. This approach is therefore usually restricted to cases where the Fourier series contains relatively few significant terms.

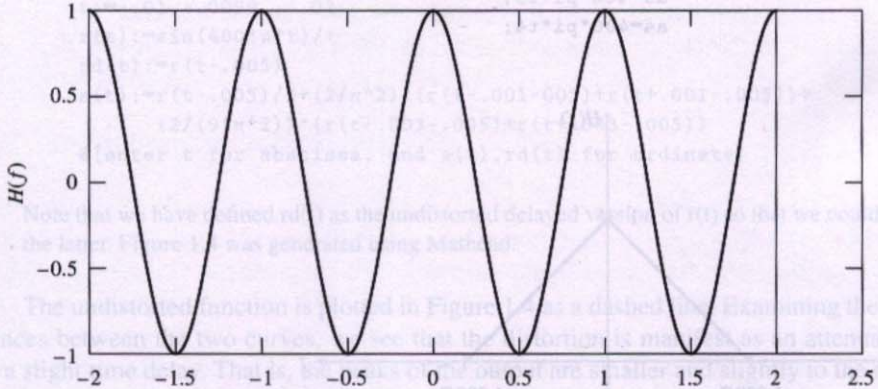


Figure 1.2 Computer plot of $H(f)$.

Example 1.1

Consider the triangular filter characteristic shown in Fig. 1.3. Assume that the phase characteristic is linear with slope $-2\pi t_0$. Find the output of this filter when the input signal is

$$r(t) = \frac{\sin 400\pi t}{t}$$

Solution: We must first expand $H(f)$ in a Fourier series to get

$$H(f) = \frac{1}{2} + \frac{4}{\pi^2}\cos\frac{\pi f}{1000} + \frac{4}{9\pi^2}\cos\frac{2\pi f}{1000} + \frac{4}{25\pi^2}\cos\frac{5\pi f}{1000} + \cdots$$

The signal $r(t)$ is bandlimited, so that all frequencies are passed by the filter. This is true because $R(f)$ is zero at frequencies above 200 Hz and the filter cuts off at $f = 1{,}000$ Hz. If we retain the first three nonzero terms in the series, the output becomes

$$s(t) = \frac{1}{2}r(t - t_0) + \frac{2}{\pi^2}\left[r\left(t - \frac{1}{2000} - t_0\right) + r\left(t + \frac{1}{2000} - t_0\right)\right]$$
$$+ \frac{2}{9\pi^2}\left[r\left(t - \frac{3}{2000} - t_0\right) + r\left(t + \frac{3}{2000} - t_0\right)\right]$$

This result is sketched as Fig. 1.4 for $t_0 = 0.005$ second.

Computer Exercise

Plot the result of Example 1.1 using both Mathcad and MATLAB. The MATLAB instructions are:

```
t=-.01:.0005:.02;
to=.005;
t1=t-to;
t2=t1-.001;
t3=t1+.001;
t4=t1-.003;
t5=t1+.003;
a1=400*pi*t1;
a2=400*pi*t2;
a3=400*pi*t3;
a4=400*pi*t4;
```

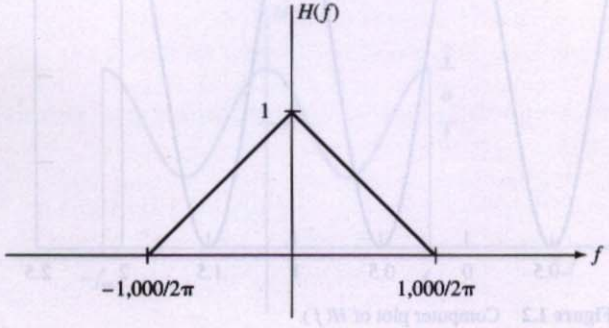


Figure 1.3 Triangular filter characteristic.

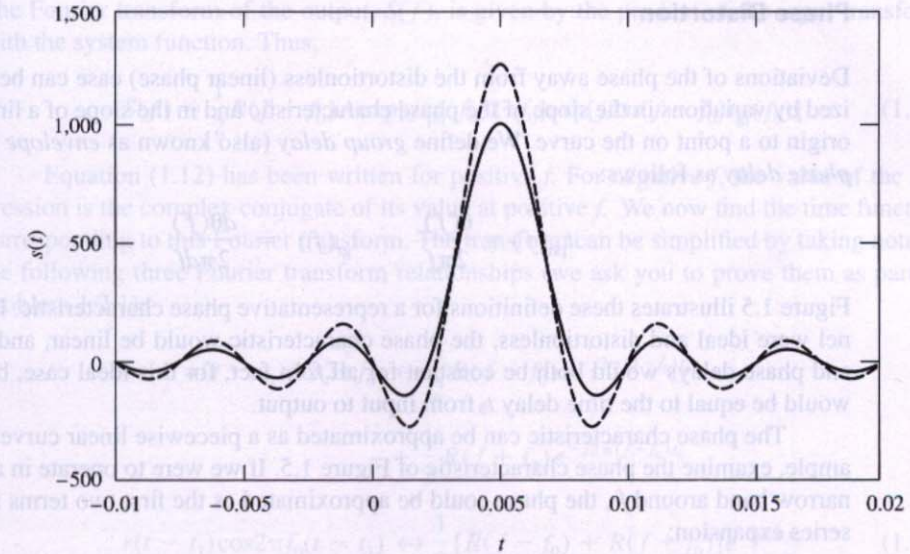


Figure 1.4 Result of Example 1.1.

```
a5=400*pi*t5;
s=sin(a1)./t1+(2/pi^2)*(sin(a2)./t2+sin(a3)./t3))
  +(2/(9*pi^2))*(sin(a4)./t4+sin(a5)./t5))
plot(t,s)
```

Note the presence of the period (.) before the division in the next-to-last statement. In MATLAB, it is important to realize that we deal with matricies. In this case, t is a (1×61) vector, so s has the same dimensions. If we attempt to execute an expression such as s=sin(t)/t, we get an error, since we are trying to divide one (1×61) vector by another (1×61) vector. To force the division to be scalar, we precede the division sign by a period.

In Mathcad, the equations can be entered directly. The instructions are:

```
t:=-.01,-.0099.. .02
r(t):=sin(400*π*t)/t
rd(t):=r(t-.005)
s(t):=r(t-.005)/2+(2/π^2)*(r(t-.001-005)+r(t+.001-.005))+
     (2/(9*π^2))*(r(t-.003-.005)+r(t+.003-.005))
@[enter t for abscissa, and s(t),rd(t) for ordinate]
```

Note that we have defined rd(t) as the undistorted delayed version of r(t) so that we could plot the latter. Figure 1.4 was generated using Mathcad.

The undistorted function is plotted in Figure 1.4 as a dashed line. Examining the differences between the two curves, we see that the distortion is manifest as an attenuation and a slight time delay. That is, the peaks of the output are smaller and slightly to the right of the corresponding peaks of the undistorted waveform.

Phase Distortion

Deviations of the phase away from the distortionless (linear phase) case can be characterized by variations in the slope of the phase characteristic and in the slope of a line from the origin to a point on the curve. We define *group delay* (also known as *envelope delay*) and *phase delay* as follows:

$$t_{\text{ph}}(f) = \frac{\theta(f)}{2\pi f} \qquad t_{\text{gr}}(f) = \frac{d\theta(f)}{2\pi df} \tag{1.9}$$

Figure 1.5 illustrates these definitions for a representative phase characteristic. If the channel were ideal and distortionless, the phase characteristic would be linear, and the group and phase delays would both be constant for all f. In fact, for this ideal case, both delays would be equal to the time delay t_0 from input to output.

The phase characteristic can be approximated as a piecewise linear curve. As an example, examine the phase characteristic of Figure 1.5. If we were to operate in a relatively narrow band around f_0, the phase could be approximated as the first two terms in a Taylor series expansion:

$$\begin{aligned}\theta(f) &\approx \theta(f_0) + \frac{d\theta(f_0)}{df}(f - f_0) \\ &= t_{\text{ph}}(f_0)f_0 + (f - f_0)t_{\text{gr}}(f_0)\end{aligned} \tag{1.10}$$

Equation (1.10) applies for positive frequency, and its negative applies for negative frequency. This is so because the phase characteristic for a real system must be an odd function of frequency.

Now suppose that the amplitude factor is constant, that is, $A(f) = A$, and a wave of the form $r(t)\cos 2\pi f_0 t$ forms the input to the system. The Fourier transform of the input is found from the modulation property of the transform (see Chapter 3):

$$r(t)\cos 2\pi f_0 t \leftrightarrow \frac{1}{2}[R(f - f_0) + R(f + f_0)] \tag{1.11}$$

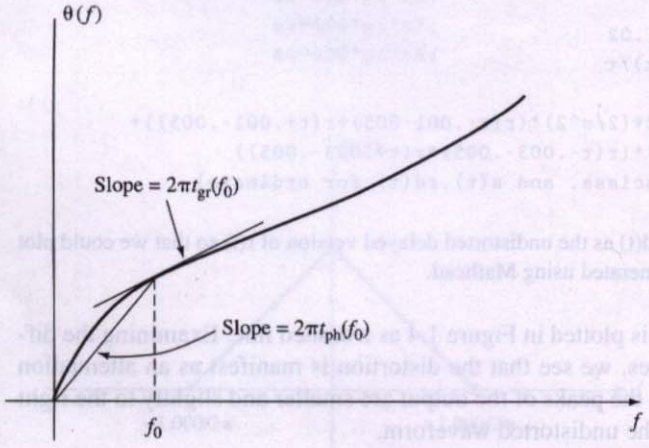


Figure 1.5 Group and phase delay.

The Fourier transform of the output, $S(f)$, is given by the product of the input transform with the system function. Thus,

$$S(f) = \frac{1}{2}R(f - f_0)A\exp[jt_{ph}(f_0)2\pi f_0]\exp[j2\pi(f - f_0)t_{gr}(f_0)] \tag{1.12}$$

Equation (1.12) has been written for positive f. For negative f, the value of the expression is the complex conjugate of its value at positive f. We now find the time function corresponding to this Fourier transform. The transform can be simplified by taking note of the following three Fourier transform relationships (we ask you to prove them as part of Problem 1.2.1):

$$r(t - t_0)\cos 2\pi f_0 t \leftrightarrow \frac{1}{2}R(f - f_0)e^{-j2\pi(f-f_0)t_0}$$

$$+ \frac{1}{2}R(f + f_0)e^{-j2\pi(f+f_0)t_0}$$

$$r(t - t_1)\cos 2\pi f_0(t - t_1) \leftrightarrow \frac{1}{2}[R(f - f_0) + R(f + f_0)]e^{-j2\pi f t_1} \tag{1.13}$$

$$r(t - t_0)\cos 2\pi f_0(t - t_1) \leftrightarrow \frac{1}{2}R(f - f_0)e^{-j2\pi(f-f_0)(t_0-t_1)}$$

$$+ \frac{1}{2}R(f + f_0)e^{-j2\pi(f+f_0)(t_0-t_1)}e^{-j2\pi f t_1}$$

Using these relationships to simplify Eq. (1.12), we find that

$$s(t) = Ar[t - t_{gr}(f_0)]\cos 2\pi f_0[t - t_{ph}(f_0)] \tag{1.14}$$

Recall that this is the output due to an input $r(t)\cos 2\pi f_0 t$. This result indicates that the time-varying amplitude of the input sinusoid is delayed by an amount equal to the group delay and the oscillating portion is delayed by an amount equal to the phase delay. Both group and phase delay are evaluated at the frequency of the sinusoid. Equation (1.14) will prove significant later. Getting ahead of the game, we will work with two types of receiver: coherent and incoherent. Incoherent receivers operate only upon the amplitude of the received signal, so the group delay is critical to the operation of the receiver. On the other hand, coherent receivers use all of the information about the waveform, so phase delay is also important.

1.2.2 Typical Communication Channels

All communication systems contain a *channel,* which is the medium that connects the receiver to the transmitter. The channel may consist of copper wires, coaxial cable, fiber optic cable, waveguide, air (including the upper atmosphere in the case of satellite transmission), or a combination of these. All channels have a maximum frequency beyond which input signal components are almost entirely attenuated. This is due to the presence of distributed capacitance and inductance. As frequencies increase, the parallel capacitance approaches a short circuit and the series inductance approaches an open circuit.

Many channels also exhibit a low-frequency cutoff due to the dual of the foregoing effects. If there is a low-frequency cutoff, the channel can be modeled as a bandpass filter. If there is no low-frequency cutoff, the channel model is a lowpass filter.

Communication channels are categorized according to bandwidth. There are three generally used grades of channel: narrowband, voiceband, and wideband.

Bandwidths up to 300 Hz are in the *narrow band;* that is, they are *telegraph grade.* They can be used for slow data transmission, on the order of 600 bits per second (*bps*). Narrowband channels cannot reliably be used for unmodified voice transmissions.

Voice-grade channels have bandwidths between 300 Hz and 4 kHz. While they were originally designed for analog voice transmission, they are regularly used to transmit data at rates on the order of 10 kilobits per second *(kbps).* Some forms of compressed video can be sent on voice-grade channels. The public telephone (subscriber loop) circuits are voiceband.

Wideband channels have bandwidths greater than 4 kHz. They can be leased from a carrier (e.g., a telephone company) and can be used for high-speed data, video, or multiple voice channels.

We now give a brief overview of the variety of communication channels in use today. We then focus on telephone channels, since their use in both analog and digital communication predominates over that of the other types of channels.

Wire, Cable, and Fiber

Copper wire, coaxial cable, or optical fibers can be used in *point-to-point* communication. That is, if we know the location of the transmitter and the location of the receiver, and if the two devices can be conveniently connected to each other, a wire connection is possible. Copper wire pairs, twisted to reduce the effects of incident noise, can be used for low-frequency communication. The bandwidth of this system is dependent upon length. The attenuation (in dB/km) follows a curve similar to that shown in Fig. 1.6.

An improvement over twisted copper pairs is realized when one moves to *coaxial cable.* The bandwidth of the channel is much higher than that of twisted wire, and multiple pairs of wires can be enclosed within a single cable sheath. The sheath which surrounds the wires shields them from incident noise, so coaxial cables can be used over longer distances than can twisted pair.

Fiber optics offers advantages over metal cable, both in bandwidth and in noise immunity. Fiber optics is particularly attractive for data communication, where the bandwidth permits much higher data rates than those achievable with metallic connectors.

Air (terrestrial) communication has both advantages and disadvantages when used as a transmission channel. The most important advantage is the ability to *broadcast* signals. You do not need to know the exact location of the receiver in order to set up a communication link. Mobile communication would not be possible without that capability. Among the disadvantages are channel characteristics that are highly dependent on frequency, additive noise, limited allocation of available frequency bands, and susceptibility to intentional interference (jamming).

Attenuation (at sea level) is a function of frequency, barometric pressure, humidity, and weather conditions. A typical curve for fair-weather conditions would resemble Fig.

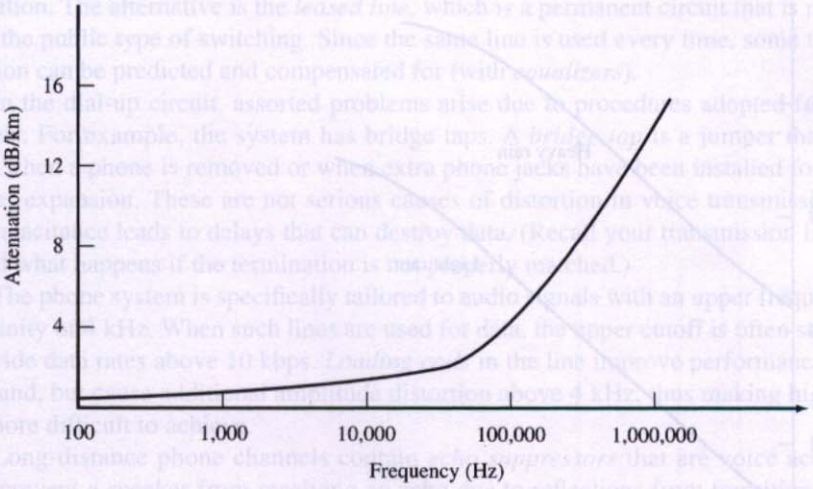


Figure 1.6 Copper wire attenuation vs. distance.

1.7(a). Visible light occupies the range of frequencies from about 40,000–75,000 GHz. Figure 1.7(b) amplifies the lower frequency portion of the attenuation curve.

Additive noise and transmission characteristics also depend upon frequency. The higher the frequency, the more the transmission takes on the characteristics of light. For example, at radio frequencies *(rf)*, in the range of 1 MHz, transmission is *not* line of sight, and reception beyond the horizon is possible. However, at *ultrahigh frequencies (uhf)*, in the range of 500 MHz and above, transmission starts acquiring some of the characteristics of light. Line of sight is needed, and humidity and obstructions degrade transmission.

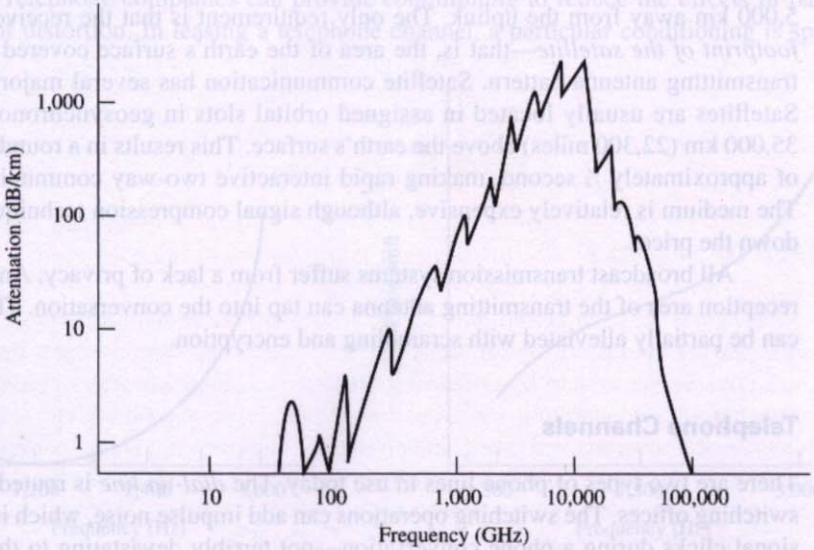


Figure 1.7(a) Attenuation vs. frequency for air.

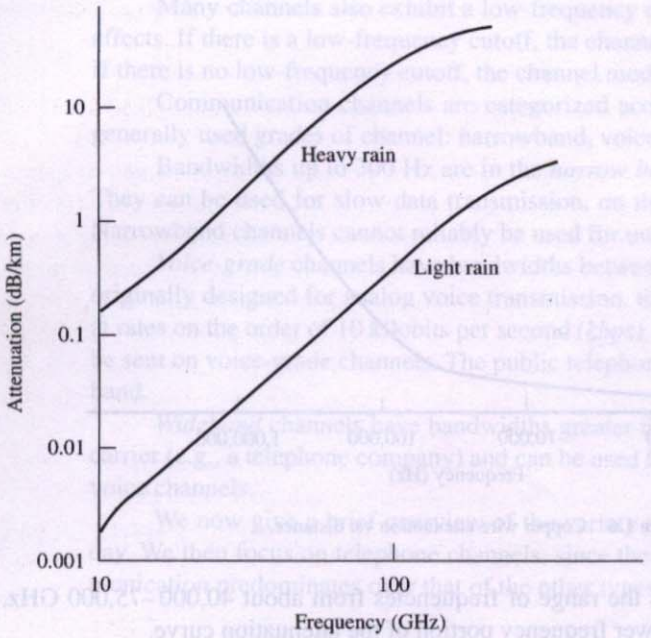


Figure 1.7(b).

At *microwave* frequencies, transmission is line of sight, and antennas must be situated in a manner to avoid obstructions.

Satellite communication provides advantages in long-distance communication. The signal is sent to the satellite via an *uplink*, and the electronics in the satellite (transponder) retransmits this signal to the *downlink*. It is just as easy for the satellite to retransmit the signal to a receiver immediately adjacent to the uplink as it is to transmit to a receiver 5,000 km away from the uplink. The only requirement is that the receiver be within the *footprint of the satellite*—that is, the area of the earth's surface covered by the satellite transmitting antenna pattern. Satellite communication has several major disadvantages. Satellites are usually located in assigned orbital slots in geosynchronous orbit, about 35,000 km (22,300 miles) above the earth's surface. This results in a round-trip travel time of approximately ½ second, making rapid interactive two-way communication difficult. The medium is relatively expensive, although signal compression techniques are bringing down the price.

All broadcast transmission systems suffer from a lack of privacy: Anyone within the reception area of the transmitting antenna can tap into the conversation. This shortcoming can be partially alleviated with scrambling and encryption.

Telephone Channels

There are two types of phone lines in use today. The *dial-up line* is routed through voice-switching offices. The switching operations can add impulse noise, which is heard as occasional clicks during a phone conversation—not terribly devastating to the conversation. But it should not be too surprising to learn that this causes serious problems in data com-

munication. The alternative is the *leased line,* which is a permanent circuit that is not subject to the public type of switching. Since the same line is used every time, some types of distortion can be predicted and compensated for (with *equalizers*).

In the dial-up circuit, assorted problems arise due to procedures adopted for voice channels. For example, the system has bridge taps. A *bridge tap* is a jumper that is installed when a phone is removed or when extra phone jacks have been installed for possible later expansion. These are not serious causes of distortion in voice transmission, but their capacitance leads to delays that can destroy data. (Recall your transmission line theory and what happens if the termination is not properly matched.)

The phone system is specifically tailored to audio signals with an upper frequency in the vicinity of 4 kHz. When such lines are used for data, the upper cutoff is often stretched to provide data rates above 10 kbps. *Loading coils* in the line improve performance in the voiceband, but cause additional amplitude distortion above 4 kHz, thus making higher bit rates more difficult to achieve.

Long-distance phone channels contain *echo suppressors* that are voice activated. These prevent a speaker from receiving an echo due to reflections from transitions in the channel. The time delay in activating the echo suppressors can make certain types of data operation impossible. Many telephone line data sets contain a provision to disable the echo suppressors using a tone of about 2,000 Hz.

Phone lines have amplitude characteristics that are not constant with frequency, and they therefore contribute to amplitude distortion. Figure 1.8 shows a typical attenuation, or loss, curve. The loss is given in decibels *(dB)* and is relative to attenuation at about 1,000 Hz, where the minimum loss occurs.

Phase distortion also occurs in the phone line. A typical phase characteristic for about 7 kilometers of phone line is shown in Fig. 1.9.

Voice-grade channels are classified according to the maximum amount of attenuation distortion and maximum envelope delay distortion within a particular frequency range. Telephone companies can provide *conditioning* to reduce the effects of particular types of distortion. In leasing a telephone channel, a particular conditioning is specified,

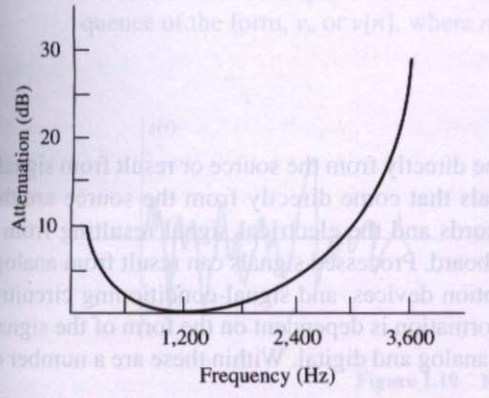


Figure 1.8 Typical telephone channel attenuation.

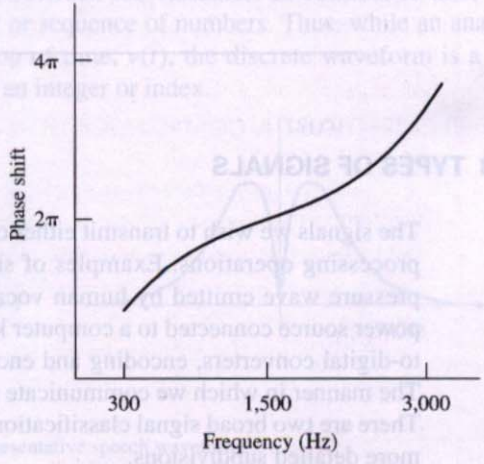


Figure 1.9 Typical telephone channel phase.

and the company guarantees a certain performance level. Naturally, the better the channel, the higher the cost will be. Table 1.1 lists typical channel conditioning characteristics for representative types of conditioning.

As an example, let us examine the C2 channel. If you were to purchase such a channel and use it within the band of frequencies between 500 Hz and 2,800 Hz, you would be guaranteed that the attenuation would not vary beyond the range of −1 dB to +3 dB (relative to response at 1,004 Hz). If you use the channel in the wider band between 300 Hz and 3,000 Hz, the guaranteed attenuation range increases to −2 dB to +6 dB. Similarly, the envelope delay variation will be less than 3,000 microseconds if you operate within the band between 500 Hz and 2,800 Hz.

In addition to the parameters given in Table 1.1, the various lines have specified losses, loss variations, maximum frequency error, and phase jitter. For example, the C1 channel specifies a loss of 16 dB ± 1 dB at 1,000 Hz. The variation in loss is limited to 4 dB over long periods of time. The frequency error is limited to 5 Hz and the phase jitter to 10°.

TABLE 1.1 TYPICAL TELEPHONE CHANNEL PARAMETERS

Conditioning	Attenuation Frequency Range	Distortion Variation	Envelope Delay Frequency Range	Distortion Variation (μs)
Basic	500–2,500	−2 to +8	800–2,600	1,750
	300–3,000	−3 to +12		
C1	1,000–2,400	−1 to +3	1,000–2,400	1,000
	300–2,700	−2 to +6	800–2,600	1,750
	300–3,000	−3 to +12		
C2	500–2,800	−1 to +3	1,000–2,600	500
	300–3,000	−2 to +6	600–2,600	1,500
			500–2,800	3,000
C4	500–3,000	−2 to +3	1,000–2,600	300
	300–3,200	−2 to +6	800–2,800	500
			600–3,000	1,500
			500–3,000	3,000

1.3 TYPES OF SIGNALS

The signals we wish to transmit either come directly from the source or result from signal-processing operations. Examples of signals that come directly from the source are the pressure wave emitted by human vocal cords and the electrical signal resulting from a power source connected to a computer keyboard. Processed signals can result from analog-to-digital converters, encoding and encryption devices, and signal-conditioning circuitry. The manner in which we communicate information is dependent on the form of the signal. There are two broad signal classifications: analog and digital. Within these are a number of more detailed subdivisions.

1.3.1 Analog Signals

An *analog signal* can be viewed as a waveform that can take on a continuum of values for any time within a range of times. Although our measuring device may be limited in resolution (e.g., it may not be possible to read an analog voltmeter more accurately than to the nearest hundredth of a volt), the actual signal can take on an infinity of possible values. For example, you might read the value of a voltage waveform at a particular time to be 13.45 volts. If the voltage is an analog signal, the actual value would be expressed as an extended decimal with an infinite number of digits to the right of the decimal point.

Just as the ordinate of the function contains an infinity of values, so does the time axis. Although we may conveniently resolve the time axis into points (e.g., every microsecond on an oscilloscope), the function has a defined value for any of the infinite number of time points between any two resolution points.

An example of an analog signal is a human speech waveform. We illustrate a representative waveform and its Fourier transform in Fig. 1.10. Note that we show only the magnitude of the Fourier transform. If the speech waveform resulted from someone whistling into a microphone, the time waveform would be a sinusoid, and the Fourier transform would be an impulse at the whistling frequency. If the person hummed into a microphone, the time waveform would be periodic with fundamental frequency equal to the frequency at which the person is humming. The Fourier transform would consist of impulses at the fundamental frequency and at its harmonics.

1.3.2 Analog Sampled Signals

Suppose that an analog time signal is defined only at discrete time points. For example, suppose you read a voltage waveform by sending values to a voltmeter every microsecond. The resulting function is known only at these discrete points in time. This results in a *discrete time function,* or a *sampled waveform.* It is distinguished from a continuous analog waveform by the manner in which we specify the function. In the case of the continuous analog waveform, we must either display the function (e.g., graphically, on an oscilloscope) or give a functional relationship between the variables. In contrast to this, the discrete signal can be thought of as a list or sequence of numbers. Thus, while an analog waveform can be expressed as a function of time, $v(t)$, the discrete waveform is a sequence of the form, v_n or $v[n]$, where n is an integer or index.

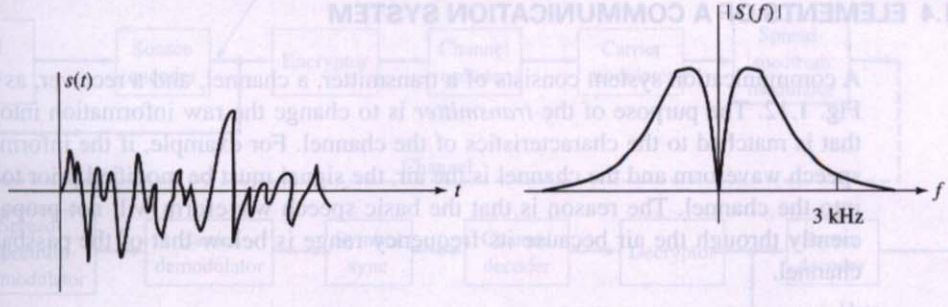


Figure 1.10 Representative speech waveform.

Discrete signals can be visualized as pulse waveforms. Figure 1.11 shows an analog time function and the resulting sampled pulse waveform. (We will refer to this sampled waveform later as *pulse amplitude modulation,* or *PAM.*)

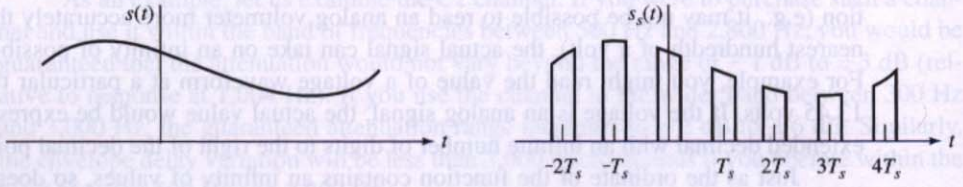


Figure 1.11 Discrete waveform derived from analog time function.

1.3.3 Digital Signals

A *digital signal* is a form of sampled or discrete waveform, but each number in the list can now take on only specific values. For example, if we were to take a sampled voltage waveform and round each value to the nearest tenth of a volt, the result would be a digital signal.

We can use a thermometer as an example of all three types of signal. If the thermometer has a dial or a tube of mercury, the output is an analog signal: We can read the temperature at any time and to any desired degree of accuracy (limited, of course, by the resolution of the reader—human or mechanical).

Suppose now that the thermometer consists of a dial, but that it is updated only once every minute. The result is an analog sampled signal.

If the transducer now takes the form of a numerical readout, the thermometer becomes digital. The readout is the result of sampling the temperature (perhaps every minute) and then displaying the sampled temperature to a predetermined resolution (perhaps the nearest tenth of a degree).

Digital signals result from many devices. For example, dialing[1] a telephone number produces 1 of 12 possible signals, depending on which button is pressed. Other examples include pressing keys on a bank automated teller machine (ATM) and using a computer keyboard. Digital signals also result from performing analog-to-digital conversion operations.

1.4 ELEMENTS OF A COMMUNICATION SYSTEM

A communication system consists of a transmitter, a channel, and a receiver, as shown in Fig. 1.12. The purpose of the *transmitter* is to change the raw information into a format that is matched to the characteristics of the channel. For example, if the information is a speech waveform and the channel is the air, the signal must be modified prior to insertion into the channel. The reason is that the basic speech waveform will not propagate efficiently through the air because its frequency range is below that of the passband of the channel.

[1]The word *dialing* is as obsolete as the word *clockwise.* It is a carryover from the early days of telephony when telephone sets contained a rotary dial used to enter numbers and control a *step-by-step* rotary switch.

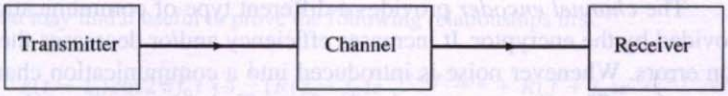


Figure 1.12 Block diagram of communication system.

The *channel* connects the transmitter to the receiver and may take the form of one of the channels described in Section 1.2.

The *receiver* accepts the signal from the channel and processes it to permit interfacing with the final destination (e.g., the human ear or a computer monitor).

Figure 1.13 shows an expansion of the simplified block diagram of a communication system. We shall briefly describe the function of each block. These functions will be expanded upon in later chapters of the text.

The *signal source* (or *transducer*) is the starting point of our system. It may be a microphone driven by a pressure wave from a human source or a musical instrument, a measuring device that is part of a monitoring system, or a data source such as a computer keyboard or a numeric pad on an ATM. Its output is a time waveform, usually electrical.

The *source encoder* operates upon one or more signals to produce an output that is compatible with the communication channel. The device could be as simple as a lowpass filter in an analog communication system, or it could be as complex as a converter that accepts analog signals and produces a periodic train of output *symbols*. These symbols may be binary (1's and 0's), or may be members of a set with more than two elements. When channels are used to communicate signals from more than one source at the same time, the source encoder contains a *multiplexer.*

With electrical communication replacing written communication, security has become increasingly important. We must assure that only the intended receiver can understand the message and that only the authorized sender can transmit it. *Encryption* provides such security. As unauthorized receivers and transmitters become more sophisticated and computers become larger and faster, the challenges of secure communication become greater. In analog systems, security is often provided using *scrambling* systems, as in pay television, and privacy devices, as with telephones.

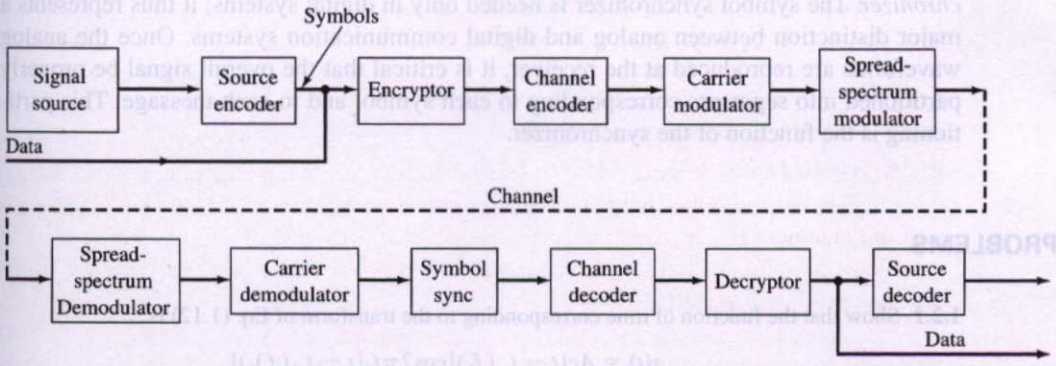


Figure 1.13 Expanded communication system block diagram.

The *channel encoder* provides a different type of communication security than that provided by the encryptor. It increases efficiency and/or decreases the effects of transmission errors. Whenever noise is introduced into a communication channel, errors occur at the receiver. These errors can take the form of changes in an analog signal, or they can make it possible for one transmitted symbol to be interpreted as a different symbol at the receiver in a digital system. We can decrease the effects of errors by providing structure to the messages in the form of redundancy. In its simplest form, this would require that the messages be repeated. We sometimes intentionally distort an analog signal to decrease the effects of frequency-sensitive noise (as, for example, in pre-emphasis/de-emphasis systems and Dolby sound systems). In digital communication, we often use *forward error correction,* in which encoding permits error correction without the necessity of the receiver asking the transmitter for additional information.

The output of the channel encoder is either a processed analog signal or a digital signal composed of symbols. For example, in a binary system, the output would be a train of 1's and 0's. We need to modify the channel encoder output signal in a manner that matches the characteristics of the channel. The *carrier modulator* produces an analog waveform that is transmitted efficiently through the system. The waveform is selected in order to provide for efficiency and also to permit multiple use of the channel by several transmitters. For analog signal sources, the carrier modulator modifies the range of signal frequencies in order to allow efficient transmission. In the case of digital signal sources, the modulator produces signal segments (bursts) corresponding to the discrete symbols at its input.

Spread spectrum is a technique for providing some immunity to frequency-selective effects such as interference and fading. A signal is spread over a wide range of frequencies so that single-tone interference affects only a small portion of the signal. Spread spectrum also has other advantages, related to simplified methods of sharing a channel among multiple users. Since an unauthorized listener may mistake a spread spectrum signal for wideband noise, the technique provides some (limited) additional security beyond that afforded by encryption.

We have been describing the blocks that form the first half of Fig. 1.13. The second half of the figure comprises the receiver, which is simply a mirror image of the transmitter. It must "undo" each operation that was performed at the transmitter. The only variation from this one-to-one correspondence is that the *carrier modulator* of the transmitter has been replaced by two blocks in the receiver: the *carrier demodulator* and the *symbol synchronizer.* The symbol synchronizer is needed only in digital systems; it thus represents a major distinction between analog and digital communication systems. Once the analog waveforms are reproduced at the receiver, it is critical that the overall signal be properly partitioned into segments corresponding to each symbol and to each message. This partitioning is the function of the synchronizer.

PROBLEMS

1.2.1 Show that the function of time corresponding to the transform of Eq. (1.12) is

$$s(t) = Ar[t - t_{gr}(f_0)]\cos 2\pi f_0[t - t_{ph}(f)_0)]$$

Hint: You may find it useful to prove the following relationships first:

$$r(t - t_0)\cos 2\pi f_0 t \leftrightarrow \frac{1}{2}[R(f - f_0)e^{-j2\pi(f-f_0)t_0} + R(f + f_0)e^{-j2\pi(f+f_0)t_0}]$$

$$r(t - t_1)\cos 2\pi f_0(t - t_1) \leftrightarrow \frac{1}{2}[R(f - f_0) + R(f + f_0)]e^{-j2\pi f t_1}$$

$$r(t - t_0)\cos 2\pi f_0(t - t_1) \leftrightarrow \frac{1}{2}[R(f - f_0)e^{-j2\pi(f-f_0)(t_0-t_1)} + R(f + f_0)e^{-j2\pi(f+f_0)(t_0-t_1)}]e^{-j2\pi f t_1}$$

1.2.2 Find the output of the filter of Fig. 1.3 when the input is

$$r(t) = \frac{\sin 200\,\pi t}{t} + \frac{5 \sin 600\,\pi t}{t}$$

1.2.3 Find the output of a typical telephone line when the input is

(a) $\cos 2\pi \times 500\,t + \cos 2\,\pi \times 1{,}000\,t$

(b) A periodic triangle of frequency 1 kHz.

1.2.4 White noise forms the input to a telephone line with amplitude and phase as shown in Figs. 2.12 and 2.13. The height of the two-sided noise power spectral density is K. Find the output power.

2

Signal Analysis

2.0 PREVIEW

What We Will Cover and Why You Should Care

By this stage of your education, you have probably heard of Fourier series expansions. You have probably also learned that the earlier part of any course of study tends to be the least interesting; certain groundwork must be laid prior to getting into the interesting applications.

The purpose of this chapter is to put the study of signals into proper perspective in the much broader area of applied mathematics. Signal analysis, and indeed most of communication theory, is a mathematical science. Probability theory and transform analysis techniques form the backbone of all communication theory. Both of these disciplines fall clearly within the realm of mathematics.

It is quite possible to study communication systems without relating the results to more general mathematical concepts, but this approach is narrow minded and tends to downgrade engineers (a phenomenon to which we need not contribute). More important, the narrow-minded approach reduces a person's ability to extend existing results to new problems.

After studying the material in this chapter, you will:

- Understand the theory of applications of Fourier series
- Be able to find the Fourier transform of various functions
- Appreciate the value of the Fourier transform
- Know the important properties of the Fourier transform

Necessary Background

The only prerequisite you need to understand the material in this chapter is basic calculus.

2.1 FOURIER SERIES

A function can be represented approximately over a given interval by a linear combination of members of an orthogonal *set* of functions. If the set of functions is denoted as $g_n(t)$, this statement can be written as

$$s(t) \approx \sum_{n=-\infty}^{\infty} c_n g_n(t) \tag{2.1}$$

An *orthogonal* set of functions is a set with the property that a particular operation performed between any two distinct members of the set yields zero. You have learned that vectors are orthogonal if they are at right angles to each other. The *dot product* of any two distinct vectors is zero. This means that one vector has nothing in common with the other. The projection of one vector onto another is zero. A function can be considered an infinite-dimensional vector (think of forming a sequence by sampling the function), so the concepts from vector spaces have direct application to function spaces. There are many possible orthogonal sets of functions, just as there are many possible orthogonal sets of three-dimensional vectors (e.g., consider any rotation of the three rectangular unit vectors).

One such set of orthogonal functions is the set of harmonically related sines and cosines. That is, the functions

$$\sin 2\pi f_0 t,\ \sin 4\pi f_0 t,\ \sin 6\pi f_0 t \ldots$$

$$\cos 2\pi f_0 t,\ \cos 4\pi f_0 t,\ \cos 6\pi f_0 t \ldots$$

form an orthogonal set for any choice of f_0. These functions are orthogonal over the interval between any starting point t_0 and $t_0 + 1/f_0$. That is,

$$\int_{t_0}^{t_0+\frac{1}{f_0}} g_n(t) g_m(t) dt = 0 \quad \textit{for all } n \neq m \tag{2.2}$$

where $g_n(t)$ is any member of the set of functions and $g_m(t)$ is any other member. This equation can be verified by a simple integration using the cosine-of-sum and cosine-of-difference trigonometric identities.

Example 2.1

Show that the set made up of the functions $\cos 2n\pi f_0 t$ and $\sin 2n\pi f_0 t$ is an orthogonal set over the interval $t_0 < t \leq t_0 + 1/f_0$ for any choice of t_0.

Solution: We must show that

$$\int_{t_0}^{t_0+\frac{1}{f_0}} g_n(t) g_m(t) dt = 0$$

for any two distinct members of the set. Three cases must be considered:

(a) Both $g_n(t)$ and $g_m(t)$ are sine waves.

(b) Both $g_n(t)$ and $g_m(t)$ are cosine waves.
(c) One of the two is a sine wave and the other a cosine wave.
Considering each of these cases in turn, we have the following:
Case (a):

$$\int_{t_0}^{t_0+\frac{1}{f_0}} \sin 2\pi n f_0 t \sin 2\pi m f_0 t \, dt = \frac{1}{2}\int_{t_0}^{t_0+\frac{1}{f_0}} \cos(n-m)2\pi f_0 t \, dt$$

$$-\frac{1}{2}\int_{t_0}^{t_0+\frac{1}{f_0}} \cos(n+m)2\pi f_0 t \, dt$$

We have used the trigonometric identity[1]

$$\sin A \sin B = \frac{1}{2}[\cos(A-B) - \cos(A+B)]$$

For $n \neq m$, both $n-m$ and $n+m$ are nonzero integers. We note that, for the function $\cos 2\pi k f_0 t$, the interval $t_0 < t \leq t_0 + 1/f_0$ represents exactly k periods. The integral of a cosine function over any whole number of periods is zero, so we have completed this case. Note that it is important that $n \neq m$, since if $n = m$, we have

$$\frac{1}{2}\int_{t_0}^{t_0+\frac{1}{f_0}} \cos(n-m)2\pi f_0 t \, dt = \frac{1}{2}\int_{t_0}^{t_0+\frac{1}{f_0}} 1 \, dt = \frac{1}{2f_0} \neq 0$$

Case (b):

$$\int_{t_0}^{t_0+\frac{1}{f_0}} \cos 2\pi n f_0 t \cos 2\pi m f_0 t \, dt = \frac{1}{2}\int_{t_0}^{t_0+\frac{1}{f_0}} [\cos(n-m)2\pi f_0 t$$

$$+\frac{1}{2}\cos(n+m)2\pi f_0 t]dt$$

Here we have used the trigonometric identity

$$\cos A \cos B = \frac{1}{2}[\cos(A+B) + \cos(A-B)]$$

This is equal to zero by the same reasoning applied to case (a).
Case (c):

$$\int_{t_0}^{t_0+\frac{1}{f_0}} \sin 2\pi n f_0 t \cos 2\pi m f_0 t \, dt = \frac{1}{2}\int_{t_0}^{t_0+\frac{1}{f_0}} \sin(n+m)2\pi f_0 t \, dt$$

$$+\frac{1}{2}\int_{t_0}^{t_0+\frac{1}{f_0}} \sin(n-m)2\pi f_0 t \, dt$$

The applicable trigonometric identity for this case is

[1]We will use three different trigonometric identities in this example. However, all necessary results could be derived from the two basic identities for $\cos(A+B)$ and $\sin(A+B)$.

$$sinAcosB = \frac{1}{2}[sin(A + B) + sin(A - B)]$$

To verify case (c), we note that

$$\int_{t_0}^{t_0+\frac{1}{f_0}} \sin k2\,\pi f_0 t\,dt = 0$$

for all integer values of k. This is true because the integral of a sine function over any whole number of periods is zero. (There is no difference between a sine function and a cosine function other than a shift.) Each term present in case (c) is therefore equal to zero.

It follows that the given set is an orthogonal set of time functions over the interval $t_0 < t \leq t_0 + 1/f_0$.

In the first sentence of this section, we used the word *approximately*. By that term, we are implying that Eq. (2.1) cannot always be made an equality. An orthogonal set of time functions is said to be a *complete set* if the approximation can in fact be made into an equality (with the word *equality* being interpreted in some special sense) through proper choice of the c_n weighting factors and for $s(t)$ being any member of a certain class of functions. The three rectangular unit vectors form a complete orthogonal set in three-dimensional space, while unit vectors in the x- and y-directions, by themselves, form an orthogonal set that is not complete.

We state without proof that the set of harmonic time functions

$$\cos 2\pi n f_0 t,\ \sin 2\pi n f_0 t$$

where n can take on any integer value between zero and infinity, is an orthogonal *complete* set in the space of time functions defined in the interval between t_0 and $t_0 + 1/f_0$. Therefore, a time function[2] can be expressed, in the interval between t_0 and $t_0 + 1/f_0$, by a linear combination of sines and cosines. In this case, the word *equality* is interpreted not as a pointwise equality, but in the sense that the *distance* between $s(t)$ and the series representation approaches zero as more and more terms are included in the sum. The distance is defined as

$$\int_{t_0}^{t_0+\frac{1}{f_0}} \left| s(t) - \sum_{n-0}^{\infty} c_n g_n(t) \right|^2 dt \tag{2.3}$$

The preceding is what we will mean when we talk of equality of two time functions. This type of equality will be sufficient for all of our applications.

For convenience, we define the period of the function as

$$T = \frac{1}{f_0} \tag{2.4}$$

Any time function $s(t)$ can then be written as

[2]In the case of Fourier series, the class of time functions is restricted to be that class which has a finite number of discontinuities and a finite number of maxima and minima in any one period. Also, the integral of the magnitude of the function over one period must exist (i.e., be finite).

$$s(t) = a_0\cos(0) + \sum_{n=1}^{\infty}[a_n\cos 2\pi n f_0 t + b_n \sin 2\pi n f_0 t] \tag{2.5}$$

for

$$t_0 < t < t_0 + T$$

An expansion of this type is known as a *Fourier series*. We note that the first term in Eq. (2.5) is simply a_0, since $\cos(0) = 1$. The proper choice of the constants a_n and b_n is indicated by the following relationships:

$$a_0 = \frac{1}{T}\int_{t_0}^{t_0+T} s(t)dt$$

$$a_n = \frac{2}{T}\int_{t_0}^{t_0+T} s(t)\cos 2\pi n f_0 t\, dt \tag{2.6}$$

$$b_n = \frac{2}{T}\int_{t_0}^{t_0+T} s(t)\sin 2\pi n f_0 t\, dt$$

The expression for a_0 in Eq. (2.6) can be derived by integrating both sides of Eq. (2.5). The expressions for a_n and b_n are derived from Eq. (2.5) by multiplying both sides by the appropriate sinusoid and integrating.

Note that a_0 is the *average* of the time function $s(t)$. It is reasonable to expect this term to appear by itself in Eq. (2.5), since the average value of the sines or cosines is zero. In any equality, the time average of the left side must equal the time average of the right side.

A more compact form of the Fourier series just described is obtained if one considers the orthogonal, complete set of complex harmonic exponentials, that is, the set made up of the time functions

$$\exp(j2\pi n f_0 t)$$

where n is any integer, positive or negative. This set is orthogonal over a period of $1/f_0$ sec. Recall that the complex exponential can be viewed as (actually, it is defined as) a vector of length 1 and angle $n2\pi f_0 t$ in the complex two-dimensional plane. Thus,

$$\exp(j2\pi n f_0 t) = \cos 2\pi n f_0 t + j\sin 2\pi n f_0 t \tag{2.7}$$

As before, the series expansion applies in the time interval between t_0 and $t_0 + 1/f_0$. Therefore, any time function $s(t)$ can be expressed as a linear combination of these exponentials in the interval between t_0 and $t_0 + T$ where ($T = 1/f_0$, as before):

$$s(t) = \sum_{n=-\infty}^{\infty} c_n e^{j2\pi n f_0 t} \tag{2.8}$$

The c_n are given by

$$c_n = \frac{1}{T}\int_{t_0}^{t_0+T} s(t)e^{-j2\pi n f_0 t}\,dt \tag{2.9}$$

This is easily verified by multiplying both sides of Eq. (2.8) by $e^{-j2\pi f_0 t}$ and integrating both sides.

The basic results are summed up in Eqs. (2.5) and (2.8): Any time function can be expressed as a weighted sum of sines and cosines or a weighted sum of complex exponentials in an interval. The rules for finding the weighting factors are given in Eqs. (2.8) and (2.9).

The right side of Eq. (2.5) represents a periodic function outside of the interval $t_0 < t < t_0 + T$. In fact, the period of the function is T. Therefore, if $s(t)$ happened to be periodic with period T, even though Eq. (2.5) was written to apply only within the interval $t_0 < t < t_0 + \mathrm{T}$, *it actually applies for all time.* (Think about it!)

In other words, if $s(t)$ is periodic, and we write a Fourier series that applies over one complete period, the series is equivalent to $s(t)$ for all time.

Example 2.2

Evaluate the trigonometric Fourier series expansion of $s(t)$ as shown in Fig. 2.1. This series must apply in the interval $-\pi/2 < t < \pi/2$.

Solution: We use the trigonometric Fourier series form with $T = \pi$ and $f_0 = 1/T = 1/\pi$. The series is therefore of the form

$$s(t) = a_0 + \sum_{n=1}^{\infty}[a_n \cos 2nt + b_n \sin 2nt]$$

where

$$a_0 = \frac{1}{\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos t\,dt = \frac{2}{\pi}$$

$$a_n = \frac{2}{\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos t \cos 2nt\,dt = \frac{2}{\pi}\left[\frac{(-1)^{n+1}}{2n-1} + \frac{(-1)^n}{2n+1}\right]$$

$$b_n = \frac{2}{\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos t \sin 2nt\,dt = 0$$

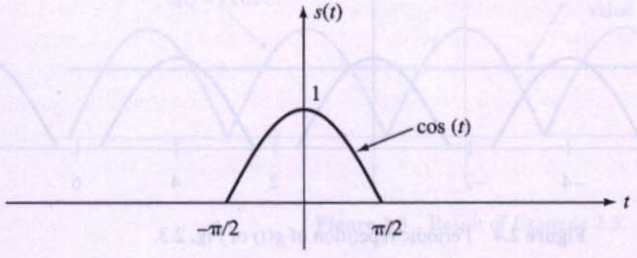


Figure 2.1 $s(t)$ for Example 2.2.

We actually did not need to evaluate the integral for b_n: Since $s(t)$ is an even function of time [i.e., $s(t) = s(-t)$], $s(t)\sin 2nt$ is an odd function, and the integral from $-T/2$ to $+T/2$ is zero. In fact, $a_n = 0$ for any odd $s(t)$. The Fourier Series is then given by

$$s(t) = \frac{2}{\pi} + \sum_{n=1}^{\infty} \frac{2}{\pi}\left[\frac{(-1)^{n+1}+}{2n-1} + \frac{(-1)^n}{2n+1}\right]\cos 2nt$$

Note that this series is also the expansion of the periodic function $s_p(t)$ shown in Fig. 2.2.

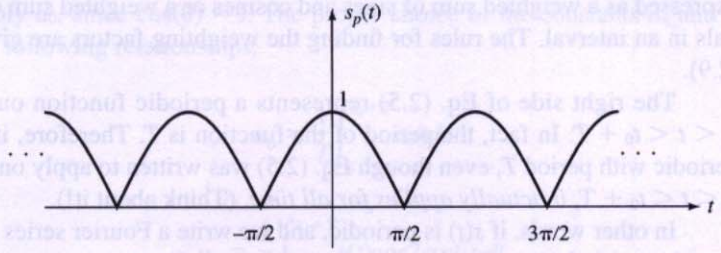


Figure 2.2 $s_p(t)$ represented by Fourier Series.

Suppose we now calculate the Fourier Series of $g(t)$ shown in Fig. 2.3, with the series required to apply in the interval $-2 < t < +2$. The result will clearly be different from that of Example 2.2. One is readily convinced of this difference once it is noted that the frequencies of the various sines and cosines will be different from those of Example 2.2. Nonetheless, for t between $-\pi/2$ and $+\pi/2$, both series expansions represent the same function of time. Both series do not, however, represent $s_p(t)$ of Fig. 2.2. The periodic function corresponding to $g(t)$, denoted as $g_p(t)$, is sketched in Fig. 2.4.

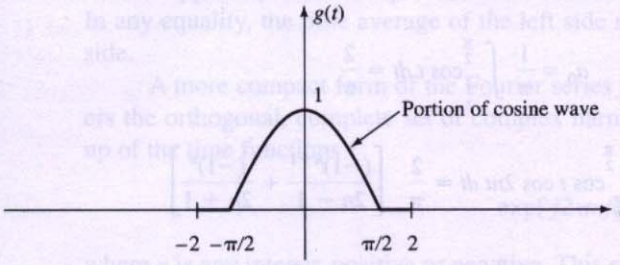


Figure 2.3 $g(t)$ similar to $s(t)$ of Example 2.2.

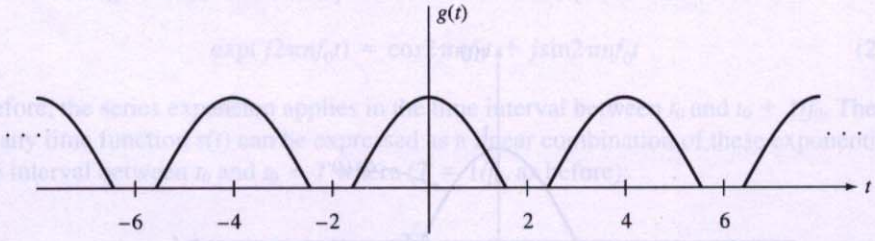


Figure 2.4 Periodic repetition of $g(t)$ of Fig. 2.3.

We conclude that the series expansion of a function in a finite interval is not unique. There are situations in which one takes advantage of this fact in order to choose the type of series that simplifies the results. (The solution of partial differential equations by separation of variables is one example.)

Example 2.3

Approximate the time function

$$s(t) = |\cos t|$$

by a constant. This constant is to be chosen so as to minimize the *error*, which is defined as the average of the square of the difference between $s(t)$ and the approximating constant. Find the "best" value of the constant.

Solution: The square error between $s(t)$ and the approximating constant C is

$$e^2(t) = [|\cos t| - C]^2$$

The average square error is found by integrating the square error over one period and dividing by the period:

$$\{e^2(t)\}_{avg} = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} [|\cos t| - C]^2 \, dt$$

We could evaluate this integral and differentiate with respect to C in order to minimize the error, or we can borrow a result from orthogonal vector spaces. We used orthogonal sets to approximate a general function. *Even if we do not use a sufficient number of terms to describe the function exactly, the weighting coefficients are chosen as if there were enough terms.* This is an important property of orthogonal sets. It is easier to visualize using vectors. If you think of approximating a three-dimensional vector with a sum of vectors in only two directions, you would not be able to make the summation identical to the vector. However, you would choose the weighting terms for the two given dimensions just as if there were sufficient terms (i.e., they would still be given by the projection of the vector on the appropriate axis). This is so because the missing vector is orthogonal to the vectors that are present and therefore has nothing in common with them. The same is true of the Fourier series: If terms are missing, the coefficients are chosen in the same manner as if all terms were present.

Accordingly, we must approximate $s(t)$ by the first term in its Fourier series expansion. The best value to choose for the constant C is the a_0, or constant, term in the Fourier series expansion. This particular expansion was evaluated in Example 2.2, where the value of a_0 was found to be $2/\pi$. The function $s(t)$ and its dc approximation are shown in Fig. 2.5.

In sum, if we wish to approximate $|\cos t|$ by a constant so as to minimize the mean square error, the best value of the constant to choose is $2/\pi$.

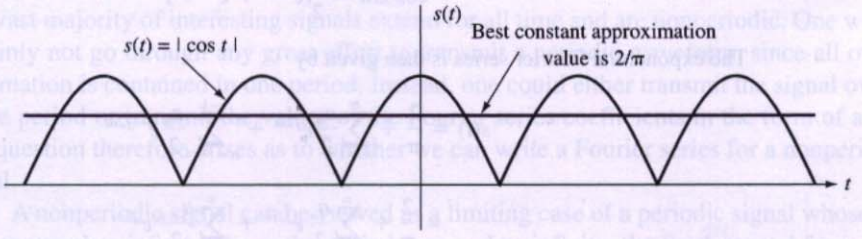


Figure 2.5 Result of Example 2.3.

2.2 COMPLEX FOURIER SPECTRUM (LINE SPECTRUM)

In finding the complex Fourier series representation of a function of time, we assign a complex weighting factor c_n to each value of n. These c_n can be plotted as a function of n. Note that this really requires two graphs, since the c_n are, in general, complex numbers. One plot can represent the magnitude of c_n and the second plot the phase. Alternatively, the real and imaginary parts could be plotted. Note further that this graph would be discrete; that is, it has nonzero value only for discrete values of the abscissa (e.g., $c_{1/2}$ has no meaning).

A more meaningful quantity than n to plot as the abscissa would be n times f_0, a quantity corresponding to the frequency of the complex exponential for which c_n is a weighting coefficient. This plot of c_n vs. nf_0 is called the *complex Fourier spectrum.*

Example 2.4

Find the complex Fourier spectrum of a full-wave rectified cosine wave. This wave is shown in Fig. 2.6 and is given by

$$s(t) = |\cos t|$$

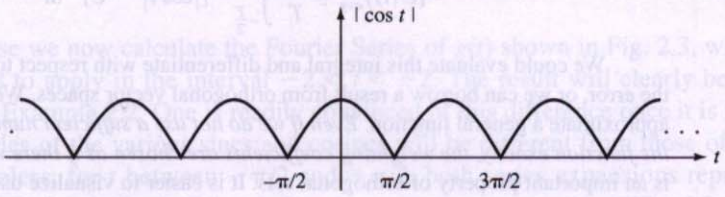


Figure 2.6 $s(t)$ for Example 2.4.

Solution: To find the complex Fourier spectrum, we must first find the exponential (complex) Fourier series expansion of the given waveform. As in Example 2.2, $f_0 = 1/\pi$. We could evaluate the c_n from Eq. (2.9) and find the Fourier series directly. However, we have already found the trigonometric Fourier series of this function in Example 2.2, namely,

$$s(t) = \frac{2}{\pi} + \sum_{n=1}^{\infty} \frac{2}{\pi}\left[\frac{(-1)^{n+1}}{2n-1} + \frac{(-1)^n}{2n+1}\right]\cos 2nt$$

We can expand the cosine function into complex exponentials by using Euler's identity. That is,

$$\cos 2nt = \frac{1}{2}[e^{j2nt} + e^{-j2nt}]$$

The exponential Fourier series is then given by

$$s(t) = \frac{2}{\pi} + \sum_{n=1}^{\infty} \frac{a_n}{2} e^{j2nt} + \sum_{n=-\infty}^{-1} \frac{a_n}{2} e^{-j2nt}$$

$$= \frac{2}{\pi} + \sum_{n=1}^{\infty} \frac{a_n}{2} e^{j2nt} + \sum_{n=1}^{\infty} \frac{a_{-n}}{2} e^{j2nt}$$

We have made a change of variables in the last summation. We see that the c_n are related to the a_n by

$$c_n = \frac{a_n}{2} \quad for\ n > 0$$

$$c_n = \frac{a_{-n}}{2} \quad for\ n < 0$$

$$c_0 = \frac{2}{\pi}$$

The resulting complex Fourier spectrum (line spectrum) is sketched in Fig. 2.7. Note that only one plot is necessary, since in this particular example, the c_n are all real numbers.

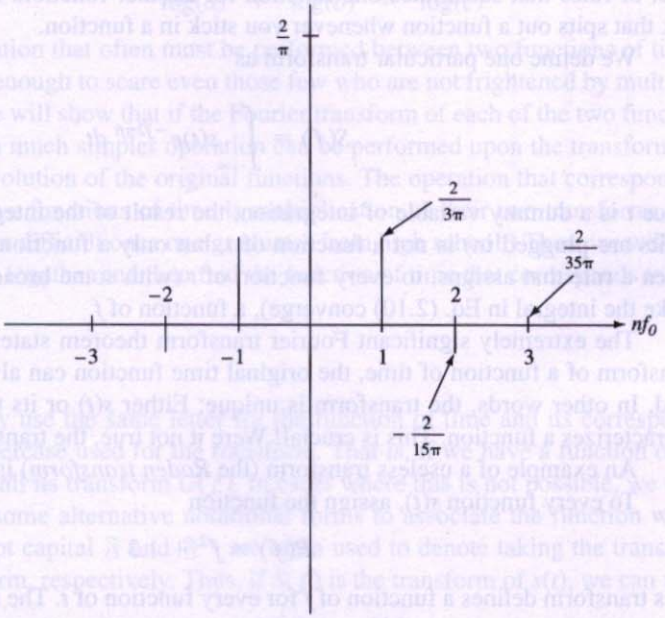


Figure 2.7 Line spectrum for Example 2.4.

2.3 FOURIER TRANSFORM

The vast majority of interesting signals extend for all time and are nonperiodic. One would certainly not go through any great effort to transmit a periodic waveform, since all of the information is contained in one period. Instead, one could either transmit the signal over a single period or transmit the values of the Fourier series coefficients in the form of a list. The question therefore arises as to whether we can write a Fourier series for a nonperiodic signal.

A nonperiodic signal can be viewed as a limiting case of a periodic signal whose period approaches infinity. Since the period approaches infinity, the fundamental frequency

f_0 approaches zero. The harmonics get closer and closer together, and in the limit, the Fourier series summation representation of $s(t)$ becomes an integral. In this manner, we could develop the Fourier integral (transform) theory.

To avoid the limiting processes required to go from Fourier series to Fourier integral, we will take an axiomatic approach. That is, we will *define* the Fourier transform and then show that the definition is extremely useful. There need be no loss in motivation by approaching the transform in this "pull out of a hat" manner, since its extreme versatility will rapidly become obvious.

What is a *transform*? Recall that a common everyday *function* is a set of rules that substitutes one number for another number. That is, $s(t)$ is a set of rules that assigns a number $s(t)$ in the *range* to any number t in the *domain*. You can think of a function as a box that spits out a number whenever you stick in a number. In a similar manner, a *transform* is a set of rules that substitutes one function for another function. It can be thought of as a box that spits out a function whenever you stick in a function.

We define one particular transform as

$$S(f) = \int_{-\infty}^{\infty} s(t)e^{-j2\pi ft}\,dt \tag{2.10}$$

Since t is a dummy variable of integration, the result of the integral evaluation (after the limits are plugged in) is not a function of t, but only a function of f. We have therefore given a rule that assigns, to every function of t (with some broad restrictions required to make the integral in Eq. (2.10) converge), a function of f.

The extremely significant Fourier transform theorem states that, given the Fourier transform of a function of time, the original time function can always be uniquely recovered. In other words, the transform is unique: Either $s(t)$ or its transform $S(f)$ uniquely characterizes a function. This is crucial! Were it not true, the transform would be useless.

An example of a useless transform (the *Roden transform*) is the following:

To every function $s(t)$, assign the function

$$R(f) = f^2 + 1.3$$

This transform defines a function of f for every function of t. The reason it has not become famous is that, among other factors, it is not unique: Given that the Roden transform of a function of time is $f^2 + 1.3$, you have not got a prayer of finding the $s(t)$ which led to that transform.

Actually, the Fourier transform theorem goes one step further than stating uniqueness: It gives the rule for recovering $s(t)$ from its Fourier transform. This rule exhibits itself as an integral and is almost of the same form as the original transform rule. That is, given $S(f)$, one can recover $s(t)$ by evaluating the integral

$$s(t) = \int_{-\infty}^{\infty} S(f)e^{j2\pi ft}\,df \tag{2.11}$$

Equation (2.11) is sometimes referred to as the *inverse transform* of S(f). It follows that this is also unique.

There are infinitely many unique transforms.[3] Why, then, has the Fourier transform achieved such widespread fame and use. Certainly, it must possess properties that make it far more useful than other transforms.

Indeed, we shall presently discover that the Fourier transform is useful in a way that is analogous to the usefulness of the common logarithm. (Remember them from high school?) In order to multiply two numbers together, we can find the logarithm of each of the numbers, add the logarithms, and then find the number corresponding to the resulting logarithm. One goes through all of this trouble in order to avoid multiplication (a frightening prospect to students). We have

$$\begin{array}{ccccc} a & \times & b & = & c \\ \Downarrow & & \Downarrow & & \Uparrow \\ \log(a) & + & \log(b) & = & \log(c) \end{array}$$

An operation that often must be performed between two functions of time is *convolution*. This is enough to scare even those few who are not frightened by multiplication! In Section 2.5, we will show that if the Fourier transform of each of the two functions of time is found first, a much simpler operation can be performed upon the transforms that corresponds to convolution of the original functions. The operation that corresponds to convolution of the two functions of time is multiplication of their two transforms. (Multiplication is no longer difficult once one graduates from high school.) Thus, we will multiply the two transforms together and then find the function of time that corresponds to the resulting transform.

Notation

We will usually use the same letter for the function of time and its corresponding transform, with uppercase used for the transform. That is, if we have a function of time called $g(t)$, we shall call its transform $G(f)$. In cases where this is not possible, we find it necessary to adopt some alternative notational forms to associate the function with its transform. The script capital $\mathfrak{F}$ and $\mathfrak{F}^{-1}$ are often used to denote taking the transform and the inverse transform, respectively. Thus, if $S(f)$ is the transform of $s(t)$, we can write

$$\mathfrak{F}[s(t)] = S(f)$$

$$\mathfrak{F}^{-1}[S(f)] = s(t)$$

A double-ended arrow is also often used to relate a function of time to its transform, the two together being known as a *transform pair*. Thus, we would write

$$s(t) \leftrightarrow S(f)$$

$$S(f) \leftrightarrow s(t)$$

[3]As two examples, consider either time scaling or multiplication by a constant. That is, define $S_1(f) = s(2f)$ or $S_2(f) = 2s(f)$. For example, if $s(t) = \sin t$, $S_1(f) = \sin 2f$ and $S_2(f) = 2\sin f$. The extension to an infinity of possible transform rules should be obvious.

2.4 SINGULARITY FUNCTIONS

We must introduce a new kind of function before proceeding to applications of Fourier theory. The new function arises whenever we analyze periodic functions. This new entity is part of the class of functions known as *singularities*. These can be thought of as derivatives of the unit step function. We begin by finding the Fourier transform of a gating function.

Example 2.5

Evaluate the Fourier transform of

$$s(t) = \begin{cases} A & |t| < \alpha \\ 0 & \text{otherwise} \end{cases}$$

Find the transform both by performing the integration and by using a computer solution. The function $s(t)$ is illustrated in Fig. 2.8.

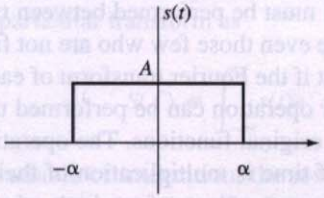


Figure 2.8 $s(t)$ for Example 2.5.

Solution: From the definition of the Fourier transform, we have

$$S(f) = \int_{-\infty}^{\infty} s(t)e^{-j2\pi ft}\,dt$$

$$= \int_{-\alpha}^{\alpha} Ae^{-j2\pi ft}\,dt = A\frac{e^{j2\pi f\alpha} - e^{-j2\pi f\alpha}}{j2\pi f}$$

$$= A\frac{\sin 2\pi f\alpha}{\pi f}$$

This transform is sketched in Fig. 2.9(a).

Note that while the Fourier transform is, in general, a complex function, the solution here turned out to be purely real. We will see the reasons for this in Section 2.6.

We now attempt to obtain the same result using computer software. Although software exists to do equations in symbolic form, we shall use Mathcad and substitute values for A and α. For purposes of illustration, we shall let $\alpha = 0.05$ and $A = 1/2\alpha$. (This should result in a Fourier transform with maximum amplitude equal to unity.)

The Mathcad instructions are as follows (the actual format for entering them depends on whether you are using Windows™ and whether you have a mouse):

$$a := .05$$

$$Fmax := \frac{2.5}{a}$$

$$A := \frac{1}{2 \cdot a}$$

$$f := 0, \frac{.01}{a} .. Fmax$$

$$F(f) := \int_{-a}^{a} A \cdot e^{(-2 \cdot j \cdot \pi \cdot f \cdot \tau)}\,d\tau$$

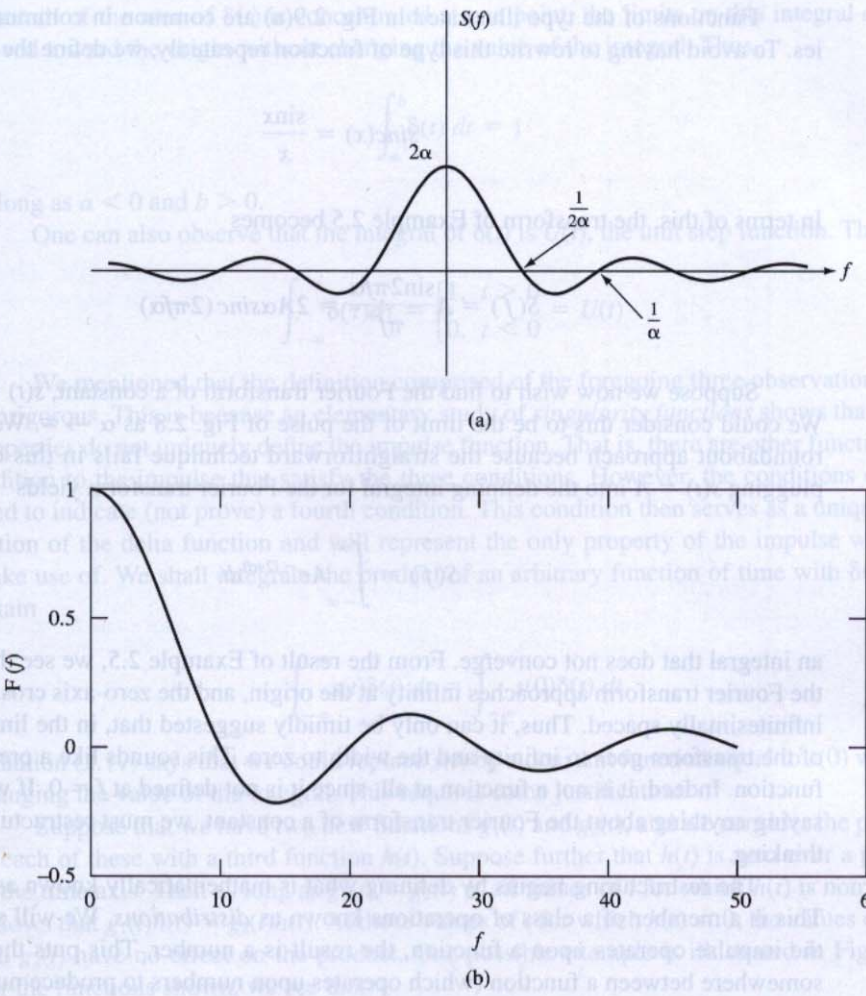


Figure 2.9 Transform of $s(t)$ for Example 2.5.

The resulting graph is shown in Fig. 2.9(b). Note that we have set Fmax so as to obtain five zeros of the function. The increment on f determines the smoothness of the curve. We have set this increment at 100 points between each pair of zeros. Of course, the smaller the increment, the longer it takes to run the simulation.

Mathcad also has a fast Fourier transform (FFT) option. There are several forms of the transform depending on whether the original data are real or complex. However, application of the FFT function requires that the number of data points be a power of 2. Since the preceding problem was so simple, and the data were given in the form of an equation, we chose to simply program the Fourier transform integral rather than use the FFT.

Functions of the type illustrated in Fig. 2.9(a) are common in communication studies. To avoid having to rewrite this type of function repeatedly, we define the *sinc* function

$$sinc(x) = \frac{\sin x}{x} \tag{2.12}$$

In terms of this, the transform of Example 2.5 becomes

$$S(f) = A\frac{\sin 2\pi f\alpha}{\pi f} = 2A\alpha sinc(2\pi f\alpha)$$

Suppose we now wish to find the Fourier transform of a constant, $s(t) = A$, for all t. We could consider this to be the limit of the pulse of Fig. 2.8 as $\alpha \to \infty$. We attempt this roundabout approach because the straightforward technique fails in this case. That is, plugging $s(t) = A$ into the defining integral for the Fourier transform yields

$$S(f) = \int_{-\infty}^{\infty} Ae^{-j2\pi ft}dt \tag{2.13}$$

an integral that does not converge. From the result of Example 2.5, we see that as $\alpha \to \infty$, the Fourier transform approaches infinity at the origin, and the zero-axis crossings become infinitesimally spaced. Thus, it can only be timidly suggested that, in the limit, the height of the transform goes to infinity and the width to zero. This sounds like a pretty ridiculous function. Indeed, it is not a function at all, since it is not defined at $f = 0$. If we insist upon saying anything about the Fourier transform of a constant, we must restructure our way of thinking.

The restructuring begins by defining what is mathematically known as the *impulse*. This is a member of a class of operations known as *distributions*. We will see that when the impulse operates upon a function, the result is a number. This puts the distribution somewhere between a function (which operates upon numbers to produce numbers) and a transform (which operates upon functions to produce functions).

We use the Greek letter *delta* (δ) to denote the impulse. While we will write $\delta(t)$ as if it were a function, we avoid difficulties by *defining* the behavior of $\delta(t)$ in all possible situations.

The usual, though nonrigorous, definition of the impulse is formed by making three simple observations. Two of these, already mentioned, are

$$\delta(t) = 0, \quad t \neq 0$$

$$\delta(t) \to \infty, \quad t = 0$$

The third property is that the total area under the impulse is equal to unity:

$$\int_{-\infty}^{\infty} \delta(t)\, dt = 1 \tag{2.14}$$

Since all of the area of $\delta(t)$ is concentrated at one point, the limits on this integral can be moved toward the origin without changing the value of the integral. Thus,

$$\int_a^b \delta(t)\,dt = 1 \tag{2.15}$$

as long as $a < 0$ and $b > 0$.

One can also observe that the integral of $\delta(t)$ is $U(t)$, the unit step function. That is,

$$\int_{-\infty}^{t} \delta(\tau)d\tau = \begin{cases} 1, & t > 0 \\ 0, & t < 0 \end{cases} = U(t) \tag{2.16}$$

We mentioned that the definition comprised of the foregoing three observations was nonrigorous. This is because an elementary study of *singularity functions* shows that these properties do not uniquely define the impulse function. That is, there are other functions in addition to the impulse that satisfy the three conditions. However, the conditions can be used to indicate (not prove) a fourth condition. This condition then serves as a unique definition of the delta function and will represent the only property of the impulse we ever make use of. We shall integrate the product of an arbitrary function of time with $\delta(t)$. We obtain

$$\int_{-\infty}^{\infty} s(t)\delta(t)\,dt = \int_{-\infty}^{\infty} s(0)\delta(t)\,dt \tag{2.17}$$

Equation (2.17) says that we could replace $s(t)$ by a constant function equal to $s(0)$ without changing the value of the integral. This requires some justification.

Suppose that we have two new functions $g_1(t)$ and $g_2(t)$, and we consider the product of each of these with a third function $h(t)$. Suppose further that $h(t)$ is zero over a portion of the time axis. Then as long as $g_1(t) = g_2(t)$ at all values of t for which $h(t)$ is nonzero, it follows that $g_1(t)h(t) = g_2(t)h(t)$. At those values of t for which $h(t) = 0$, the values of $g_1(t)$ and $g_2(t)$ have no effect on the product. One possible example is illustrated in Fig. 2.10. For the functions shown, we see that

$$g_1(t)h(t) = g_2(t)h(t) \tag{2.18}$$

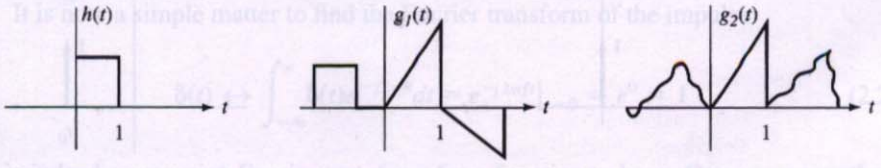


Figure 2.10 Example of $h(t)$, $g_1(t)$, and $g_2(t)$.

Returning to Eq. (2.17), we note that $\delta(t)$ is zero for all $t \neq 0$. Therefore, the product of $\delta(t)$ with any function of time depends only upon the value of that function at $t = 0$. Figure 2.11 illustrates several possible functions that have the same product with $\delta(t)$ as does $s(t)$.

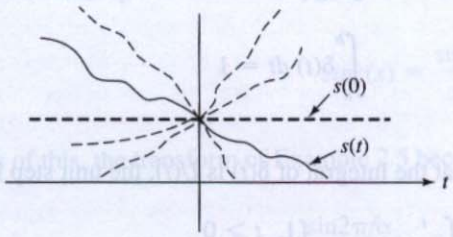


Figure 2.11 Functions having same product with $\delta(t)$.

Out of the infinity of possibilities, the constant function of time is a wise choice, since we can factor it out of the integral to get

$$\int_{-\infty}^{\infty} s(t)\delta(t)\,dt = s(0)\int_{-\infty}^{\infty} \delta(t)\,dt = s(0) \tag{2.19}$$

This is a significant result, and we will refer to it as the *sampling property* of the impulse. Note that a great deal of information about $s(t)$ has been lost, since the result depends only upon the value of $s(t)$ at one point.

A change of variables yields a shifted impulse with the analogous sampling property:

$$\int_{-\infty}^{\infty} s(t)\delta(t - t_0)\,dt = \int_{-\infty}^{\infty} s(k + t_0)\delta(k)\,dk = s(t_0) \tag{2.20}$$

Figure 2.12 shows $\delta(t)$ and $\delta(t - t_0)$. The upward-pointing arrow is a generally accepted technique to indicate an actual value of infinity. The number next to the arrow indicates the total area under the impulse, known as the *strength*. In sketching the impulse, the height of the arrow is made equal to the strength of the impulse.

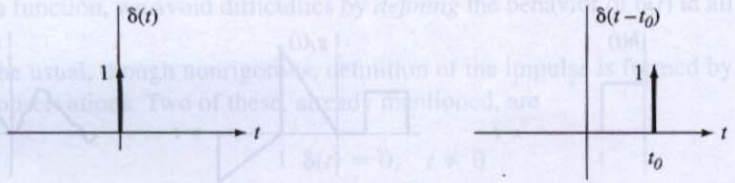


Figure 2.12 Pictorial representation of impulse.

Equations (2.19) and (2.20) are the only things one must know about the impulse. Indeed, either of them can be treated as the definition of the impulse.

Example 2.6

Evaluate the following integrals:

$$\text{(a)} \int_{-\infty}^{\infty} \delta(t)(t^2 + 1)\, dt$$

$$\text{(b)} \int_{-1}^{2} \delta(t - 1)(t^2 + 1)\, dt$$

$$\text{(c)} \int_{3}^{5} \delta(t - 1)(t^3 + 4t + 2)\, dt$$

$$\text{(d)} \int_{-\infty}^{\infty} \delta(1 - t)(t^4 + 2)\, dt$$

Solution: (a) Straightforward application of the sampling property yields

$$\int_{-\infty}^{\infty} \delta(t)(t^2 + 1)dt = 0^2 + 1 = 1$$

(b) Since the impulse falls within the range of integration,

$$\int_{-1}^{2} \delta(t - 1)(t^2 + 1)dt = 1^2 + 1 = 2$$

(c) The impulse occurs at $t = 1$, which is outside the range of integration. Therefore,

$$\int_{3}^{5} \delta(t - 1)(t^3 + 4t + 2)dt = 0$$

(d) $\delta(1 - t)$ falls at $t = 1$, since this is the value of t that makes the argument equal to zero. Therefore,[4]

$$\int_{-\infty}^{\infty} \delta(1 - t)(t^4 + 2)dt = 1^4 + 2 = 3$$

It is now a simple matter to find the Fourier transform of the impulse:

$$\delta(t) \leftrightarrow \int_{-\infty}^{\infty} \delta(t)e^{-j2\pi ft}dt = e^{-j2\pi ft}\Big|_{t=0} = e^0 = 1 \tag{2.21}$$

This is indeed a very nice Fourier transform for a function to have. One can guess that it will prove significant, since it is the unity, or identity, multiplicative element. That is, anything multiplied by 1 is left unchanged.

[4]Many arguments can be advanced for regarding the impulse as an even function. For one thing, it can be considered the derivative of an odd function. For another, the fact that the transform is real and even indicates that the impulse is an even time function, as will be explored in Section 2.6.

It is almost time to apply all of the theory we have been laboriously developing to a practical problem. Before doing that, we need simply evaluate several transforms that involve impulses.

Let us return to the evaluation of the transform of a constant, $s(t) = A$. We observed earlier that the defining integral

$$A \leftrightarrow \int_{-\infty}^{\infty} Ae^{-j2\pi ft}\,dt \tag{2.22}$$

does not converge. For $f \neq 0$, this integral is bounded by $A/\pi f$. For $f = 0$, the integral "blows up."

Since the integral defining the Fourier transform and that used to evaluate the inverse transform are quite similar, one might guess that the transform of a constant is an impulse. That is, since an impulse transforms to a constant, a constant should transform to an impulse. In the hope that this guess is valid, let us find the inverse transform of an impulse. We have

$$\delta(f) \leftrightarrow \int_{-\infty}^{\infty} \delta(f)e^{j2\pi ft}df = 1 \tag{2.23}$$

Our guess was correct: The inverse transform of $\delta(f)$ is a constant. Therefore, by applying a scaling factor, we have

$$A \leftrightarrow A\delta(f) \tag{2.24}$$

If we take the inverse transform of a shifted impulse, we develop the additional transform pair,

$$Ae^{j2\pi f_0 t} \leftrightarrow A\delta(f - f_0) \tag{2.25}$$

This guess-and-check technique deserves some comment. We stressed earlier that the uniqueness of the Fourier transform is extremely significant. That is, given $S(f)$, $s(t)$ can be uniquely recovered. Therefore, the guess-and-check technique is a perfectly rigorous one to use to find the Fourier transform of a function of time. If we can somehow guess at an $S(f)$ that yields $s(t)$ when $S(f)$ is plugged into the inversion integral, we have found the one and only transform of $s(t)$. As in the preceding example, this technique is very useful if the transform integral cannot be easily evaluated, whereas the inverse transform integral can.

Example 2.7

Find the Fourier transform of $s(t) = \cos 2\pi f_0 t$.

Solution: We make use of Euler's identity to express the cosine function in the form

$$\cos 2\pi f_0 t = \frac{1}{2}e^{j2\pi f_0 t} + \frac{1}{2}e^{-j2\pi f_0 t}$$

The Fourier transform of the cosine is then the sum of the transforms of the two exponentials, which we found in Eq. (2.25). Therefore,

$$\cos 2\pi f_0 t \leftrightarrow \frac{1}{2}\delta(f - f_0) + \frac{1}{2}\delta(f + f_0)$$

This transform is sketched in Fig. 2.13.

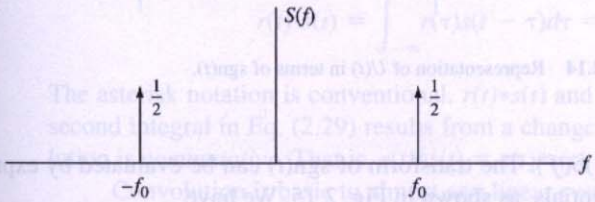


Figure 2.13 Fourier transform of cosine wave.

We can now reveal a deception we have been guilty of: Although the Fourier transform is specified by a strictly mathematical definition, and f is just an independent functional variable, we have slowly tried to brainwash you into thinking of this variable as *frequency*. Indeed, the choice of f for the symbol of the variable brings the word *frequency* to mind. We have seen several Fourier transforms that are not identically zero for negative values of f. In fact, we shall see in Section 2.6 that the transform *cannot* be zero for negative f in the case of real functions of time. Since the definition of frequency (repetition rate) has no meaning for negative values, we could never be completely correct in calling f a frequency variable.

Suppose we view the positive f-axis only. For this region, the transform of $\cos 2\pi f_0 t$ is nonzero only at the point $f = f_0$. (See Example 2.7.) Probably the only time in your earlier education that you experienced the definition of frequency is the case where the function of time is a pure sinusoid. For this function, the positive f-axis appears to have meaning when interpreted as frequency, so we shall consider ourselves justified in calling f a frequency variable.

Another transform pair that will be needed in our later work is that of a unit step function and its transform. Here, as in the case of a constant, if one simply plugs the function of time into the transform definition, the resulting integral does not converge. We could again attempt the guess-and-check technique, but due in part to the discontinuity of the step function, the technique becomes not very hopeful. The transform is relatively easy to evaluate once one realizes that

$$U(t) = \frac{1 + \text{sgn}\,(t)}{2}$$

where sgn is the *sign* function, defined by

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

$U(t)$ is illustrated in Fig. 2.14.

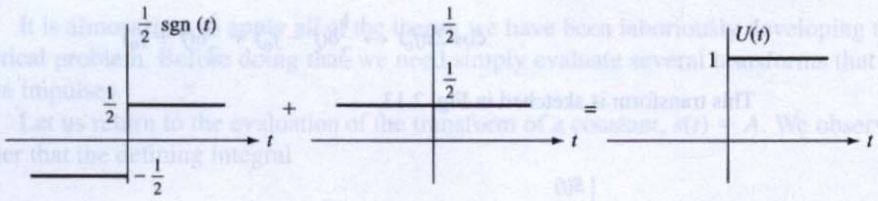


Figure 2.14 Representation of $U(t)$ in terms of sgn(t).

The transform of $\frac{1}{2}$ is $(\frac{1}{2})\delta(f)$. The transform of sgn(t) can be evaluated by expressing sgn(t) as a limit of exponentials, as shown in Fig. 2.15. We have

$$\text{sgn}\,(t) = \lim_{a\to 0}\,[e^{-a|t|}\text{sgn}\,(t)]$$

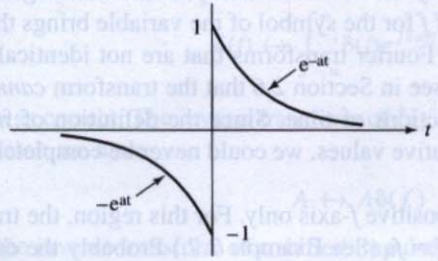


Figure 2.15 sgn(t) as a limit of exponentials.

Assuming that the order of taking the limit and taking the transform can be interchanged (generally, if the result is bounded, we can do this), we obtain

$$\begin{aligned}\mathcal{F}[\text{sgn}\,(t)] &= \lim_{a\to 0}\,\mathcal{F}[e^{-a|t|}\text{sgn}\,(t)] \\ &= \lim_{a\to 0}\left[\frac{1}{j2\pi f + a} + \frac{1}{j2\pi f - a}\right] = \frac{1}{j\pi f}\end{aligned} \tag{2.27}$$

The transform of the unit step is then given by

$$U(t) \leftrightarrow \frac{1}{j2\pi f} + \frac{1}{2}\delta(f) \tag{2.28}$$

If you have been exposed to Laplace transforms, you may recall that the Laplace transform of a unit step is $1/s$. At first glance, it appears that the Fourier transform of any function that is zero for negative t should be the same as the one-sided Laplace transform, with s replaced by $j2\pi f$. However, we see that in the case of $s(t) = U(t)$, the two transforms differ by a very important factor, $(\frac{1}{2})\delta(f)$. The explanation of this apparent discrepancy requires a study of convergence in the complex s-plane.

2.5 CONVOLUTION

We are now ready to investigate the "scary" operation referred to at the end of Section 2.3. The *convolution* of two time functions $r(t)$ and $s(t)$ is defined by the integral operation

$$r(t)*s(t) = \int_{-\infty}^{\infty} r(\tau)s(t-\tau)d\tau = \int_{-\infty}^{\infty} s(\tau)r(t-\tau)d\tau \tag{2.29}$$

The asterisk notation is conventional, $r(t)*s(t)$ and is read "$r(t)$ convolved with $s(t)$." The second integral in Eq. (2.29) results from a change of variables, and it proves that convolution is *commutative*. That is, $r(t)*s(t) = s(t)*r(t)$.

Convolution is basic to almost any linear system.

Note that the convolution of two functions of t is itself a function of t, since τ is a dummy variable of integration. The integral of Eq. (2.29) is, in general, very difficult to evaluate in closed form, as is demonstrated in the following example.

Example 2.8

Evaluate the convolution of $r(t)$ with $s(t)$, where $r(t)$ and $s(t)$ are the square pulses shown in Fig. 2.16.

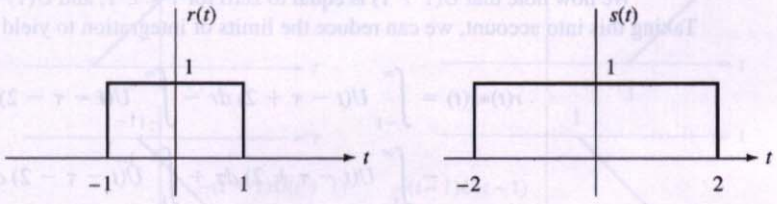


Figure 2.16 Functions for Example 2.8.

Solution: We note that the functions can be written in the form

$$r(t) = U(t+1) - U(t-1)$$

$$s(t) = U(t+2) - U(t-2)$$

where $U(t)$ is the unit step function defined by

$$U(t) = \begin{cases} 1, & t > 0 \\ 0, & t < 0 \end{cases}$$

The convolution is defined by

$$r(t)*s(t) \triangleq \int_{-\infty}^{\infty} r(\tau)s(t-\tau)\,d\tau$$

We see that

$$r(\tau) = U(\tau+1) - U(\tau-1)$$

and

$$s(t-\tau) = U(t-\tau+2) - U(t-\tau-2)$$

$$r(\tau)s(t-\tau) = U(\tau+1)U(t-\tau+2) - U(\tau+1)U(t-\tau-2)$$

$$- U(\tau-1)U(t-\tau+2) + U(\tau-1)U(t-\tau-2)$$

Therefore, breaking the integral into parts, we have

$$r(t)*s(t) = \int_{-\infty}^{\infty} U(\tau+1)U(t-\tau+2)\,d\tau$$

$$- \int_{-\infty}^{\infty} U(\tau+1)U(t-\tau-2)\,d\tau$$

$$- \int_{-\infty}^{\infty} U(\tau-1)U(t-\tau+2)\,d\tau$$

$$+ \int_{-\infty}^{\infty} U(\tau-1)U(t-\tau-2)\,d\tau$$

We now note that $U(\tau+1)$ is equal to zero for $\tau < -1$, and $U(\tau-1)$ is zero for $\tau < 1$. Taking this into account, we can reduce the limits of integration to yield

$$r(t)*s(t) = \int_{-1}^{\infty} U(t-\tau+2)\,dr - \int_{-1}^{\infty} U(t-\tau-2)\,d\tau$$

$$- \int_{1}^{\infty} U(t-\tau+2)\,d\tau + \int_{1}^{\infty} U(t-\tau-2)\,d\tau$$

To derive this, we have replaced one of the step functions by its value, unity, in the range in which the substitution applies. We now try to evaluate each integral separately. Note that

$$U(t-\tau+2) = 0, \quad \tau > t+2$$

and

$$U(t-\tau-2) = 0, \quad \tau > t-2$$

Using these facts, we have

$$\int_{-1}^{\infty} U(t-\tau+2)\,d\tau = \int_{-1}^{t+2} d\tau = t+3$$

provided that $t+2 > -1$, or equivalently, $t > -3$. Otherwise, the integral evaluates to zero. Likewise, if $t-2 > -1$, that is, $t > 1$, then

$$\int_{-1}^{\infty} U(t-\tau-2)\,d\tau = \int_{-1}^{t-2} d\tau = t-1$$

If $t + 2 > 1$, that is, $t > -1$, then

$$\int_1^{\infty} U(t - \tau + 2)\, dr = \int_1^{t+2} d\tau = t + 1$$

If $t - 2 > 1$, that is, $t > 3$, then

$$\int_1^{\infty} U(t - \tau - 2)\, d\tau = \int_1^{t-2} d\tau = t - 3$$

Using these four results, we find that

$$r(t)*s(t) = (t + 3)U(t + 3) - (t - 1)U(t - 1) - (t + 1)U(t + 1) + (t - 3)U(t - 3)$$

The four terms on the right-hand side, together with their sum, are sketched in Fig. 2.17. From this modest example, we can see that, if either $r(t)$ or $s(t)$ contains step functions, the evaluation of the convolution becomes quite involved.

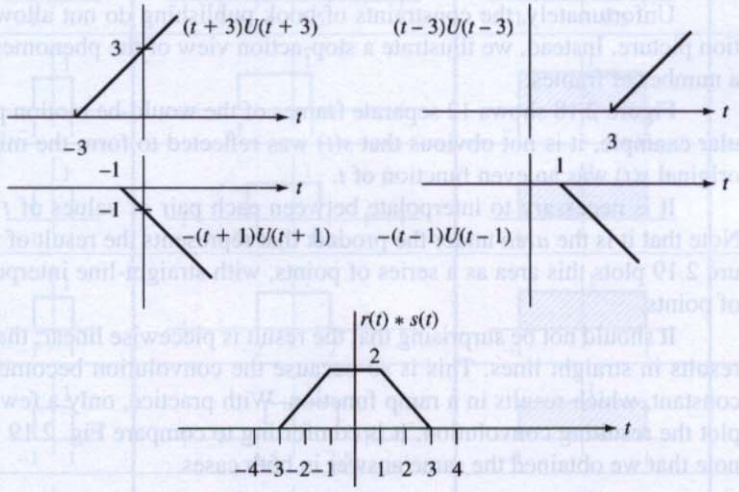


Figure 2.17 Result of Example 2.8.

2.5.1 Graphical Convolution

We claim that, for simple $r(t)$ and $s(t)$ (what we mean by *simple* should be clear at the end of this section), the result of the convolution can be obtained almost by inspection. Even in cases where $r(t)$ and $s(t)$ are quite complicated or the waveshapes are not precisely known, certain observations can be made about the convolution without actually performing the detailed integration. In many communication applications, these general observations will be sufficient, and the exact convolution will not be required.

The inspection procedure is known as *graphical convolution*. We will arrive at the technique by examining the definition of convolution. We repeat the left-hand equality of Eq. (2.29):

$$r(t)*s(t) = \int_{-\infty}^{\infty} r(\tau)s(t-\tau)d\tau$$

One of the original functions is $r(\tau)$, where the independent variable is now called τ.

The mirror image of $s(\tau)$ is represented by $s(-\tau)$, that is, $s(\tau)$ reflected around the vertical axis.

The convolution equation now tells us that for a given t, we form $s(t-\tau)$, which represents the function $s(-\tau)$ shifted to the right by t. We then take the product

$$r(\tau)s(t-\tau)$$

and integrate it (i.e., find the area under it) in order to find the value of the convolution for that particular value of t. The procedure is illustrated in Fig. 2.18 for the two functions of Fig. 2.16 from Example 2.8. The ideal way to demonstrate graphical convolution is with an animated motion picture that shows the two functions, one of which is moving across the τ-axis. The motion picture would show the two functions overlapping by varying amounts as t changes.

Unfortunately, the constraints of book publishing do not allow us to present a motion picture. Instead, we illustrate a stop-action view of the phenomenon; that is, we show a number of frames.

Figure 2.18 shows 12 separate frames of the would-be motion picture. In this particular example, it is not obvious that $s(t)$ was reflected to form the mirror image, since the original $s(t)$ was an even function of t.

It is necessary to interpolate between each pair of values of t shown in the figure. Note that it is the *area* under the product that represents the result of the convolution. Figure 2.19 plots this area as a series of points, with straight-line interpolation between pairs of points.

It should not be surprising that the result is piecewise linear; that is, the interpolation results in straight lines. This is so because the convolution becomes an integration of a constant, which results in a ramp function. With practice, only a few points are needed to plot the resulting convolution. It is comforting to compare Fig. 2.19 with Fig. 2.17 and to note that we obtained the same answer in both cases.

Example 2.9

Find the convolution of $r(t)$ with itself, where the only information given about $r(t)$ is that it is zero for $|t| > a$. That is, $r(t)$ is limited to the range between $t = -\alpha$ and $t = +\alpha$. A representative $r(t)$ is sketched in Fig. 2.20.

Solution: Not knowing $r(t)$ exactly, we certainly cannot find the resulting convolution, $r(t)*r(t)$, exactly. To see how much information we can obtain concerning this convolution, we attempt graphical convolution. Figure 2.21 is a sketch of $r(\tau)$ and $r(t-\tau)$.

We note that as t increases from zero, the two functions illustrated have less and less of the τ-axis in common. When t reaches $+2\alpha$, the two functions separate; that is, at $t = 2\alpha$, we have the situation sketched in Fig. 2.22. The two functions continue to have nothing in common for all t greater than 2α. Likewise, for negative t, one can see that the product of the two functions is zero as long as $t < -2\alpha$.

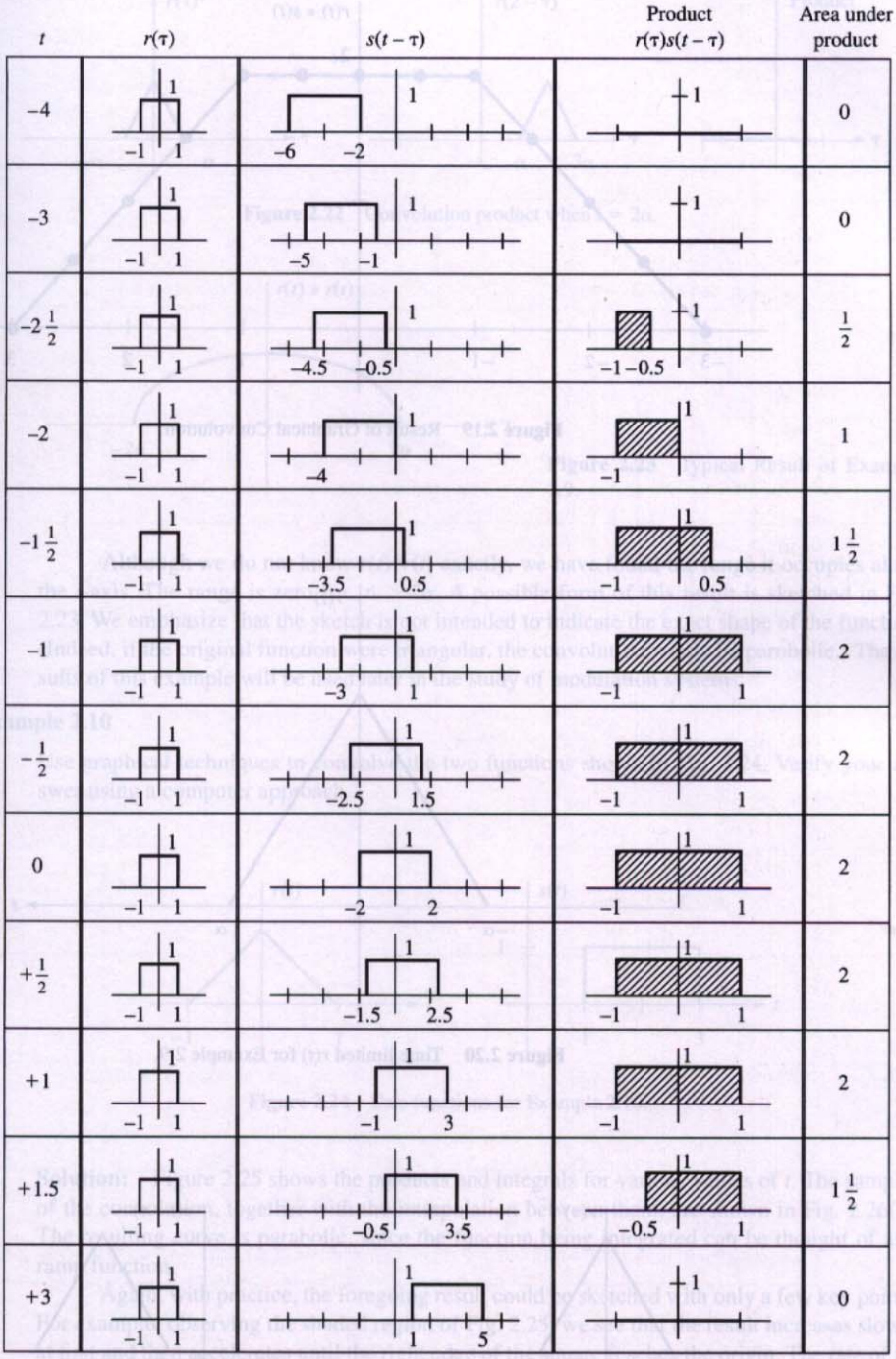


Figure 2.18 Graphical Convolution.

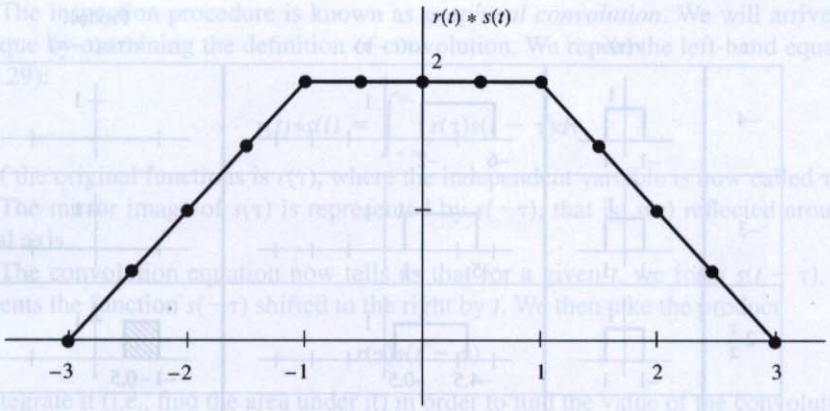


Figure 2.19 Result of Graphical Convolution.

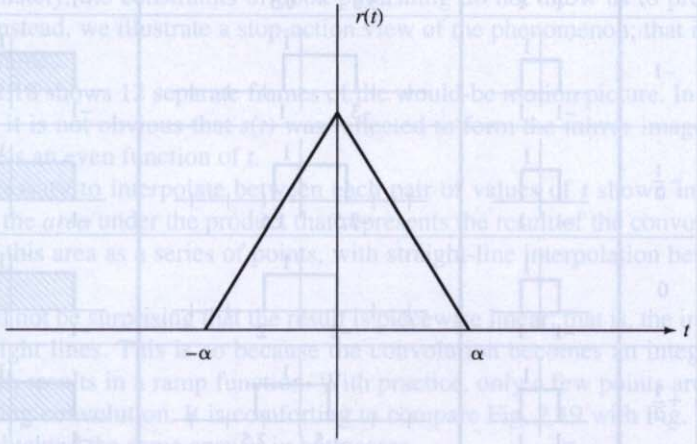


Figure 2.20 Time limited $r(t)$ for Example 2.9.

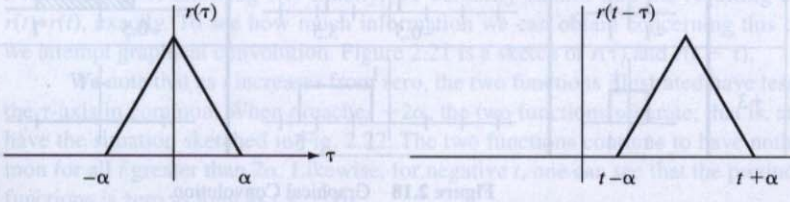


Figure 2.21 $r(\tau)$ and $r(t - \tau)$ for Example 2.9.

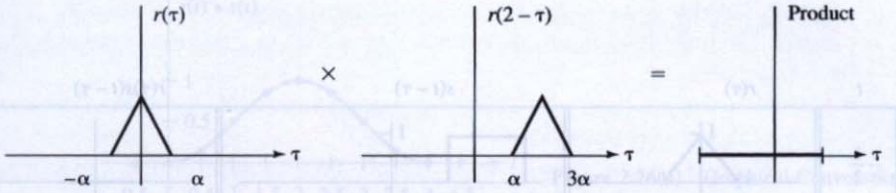


Figure 2.22 Convolution product when t = 2α.

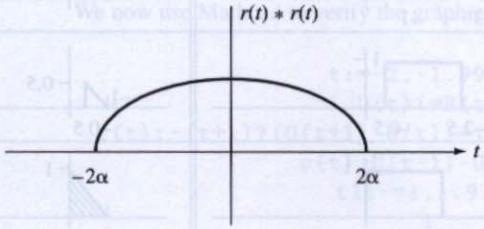


Figure 2.23 Typical Result of Example 2.9.

Although we do not know $r(t)*r(t)$ exactly, we have found the range it occupies along the t-axis. The range is zero for $|t| > 2\alpha$. A possible form of this result is sketched in Fig. 2.23. We emphasize that the sketch is not intended to indicate the exact shape of the function. (Indeed, if the original function were triangular, the convolution would be parobolic.) The results of this example will be used later in the study of modulation systems.

Example 2.10

Use graphical techniques to convolve the two functions shown in Fig. 2.24. Verify your answer using a computer approach.

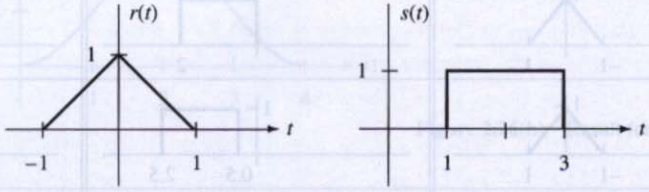


Figure 2.24 Two functions for Example 2.10.

Solution: Figure 2.25 shows the products and integrals for various values of t. The samples of the convolution, together with the interpolation between them, are shown in Fig. 2.26(a). The resulting curve is parabolic, since the function being integrated can be thought of as a ramp function.

Again, with practice, the foregoing result could be sketched with only a few key points. For example, observing the shaded region of Fig. 2.25, we see that the result increases slowly at first and then accelerates until the right edge of the square reaches the origin. The rate of increase then gets smaller again, as each incremental move to the right adds a shorter and shorter strip to the product.

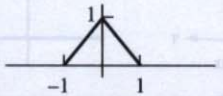
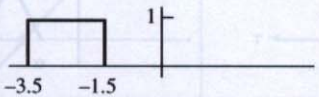
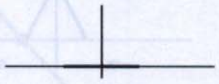
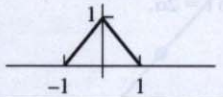
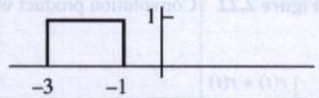
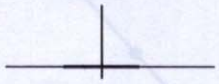
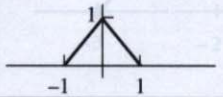
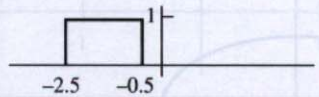
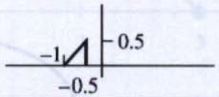
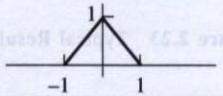
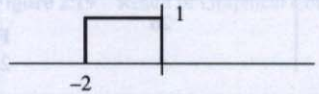
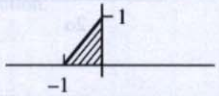
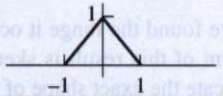
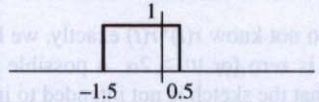
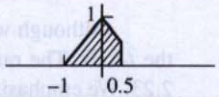
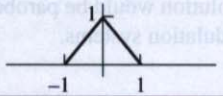
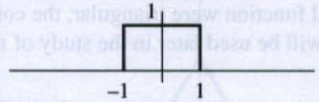
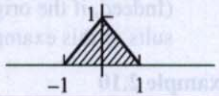
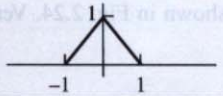
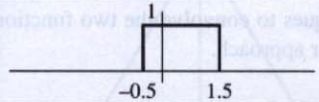
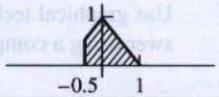
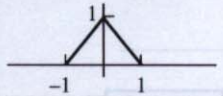
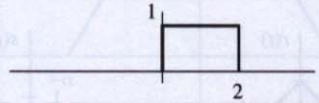
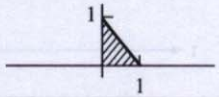
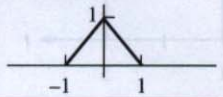
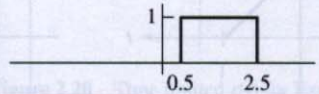
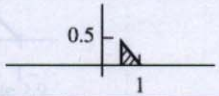
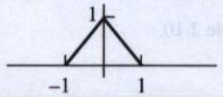
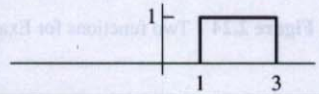
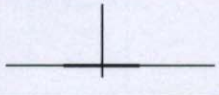
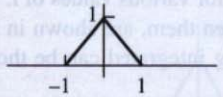
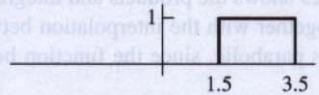
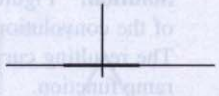
t	$r(\tau)$	$s(t-\tau)$	$r(\tau)s(t-\tau)$	Area under product
$-\frac{1}{2}$				0
0				0
$\frac{1}{2}$				$\frac{1}{8}$
1				$\frac{1}{2}$
1.5				$\frac{7}{8}$
2				1
2.5				$\frac{7}{8}$
3				$\frac{1}{2}$
3.5				$\frac{1}{8}$
4				0
4.5				0

Figure 2.25

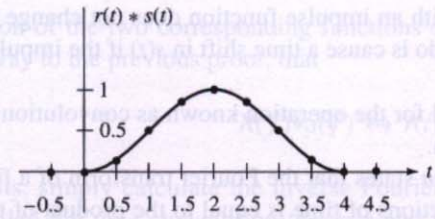


Figure 2.26(a) Graphical Convolution for Example 2.10.

We now use Mathcad to verify the graphical solution. The input code is as follows:

```
t:=-2,-1.99..2
U(t):=F(t)
r(t):-(t+1)?(U(t+1)-U(t))+(-t+1)?(U(t)-U(t-1))
s(t):U(t-1)-U(t-3)
t1:-=1,-.9..5
1
c(t1):=er(tau)?s(t1-tau)dtau
-1
```

Note that r(t) is the triangular pulse defined using gated ramps and s(t) is the square pulse. F(t) is the Mathcad expression for the unit step. The resulting convolution, c(t1), is shown in Figure 2.26(b). Note that it matches the result we obtained using graphical convolution.

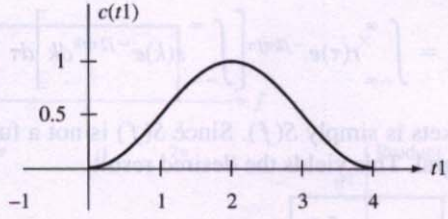


Figure 2.26(b) Result for Example 2.10.

We now investigate the operation of convolving an arbitrary time function with $\delta(t)$:

$$\delta(t)*s(t) = \int_{-\infty}^{\infty} \delta(\tau)s(t-\tau)d\tau = s(t-0) = s(t) \tag{2.30}$$

This shows that any function convolved with an impulse remains unchanged.

If we convolve $s(t)$ with the shifted impulse, $\delta(t-t_0)$, we obtain

$$\delta(t-t_0)*s(t) = \int_{-\infty}^{\infty} \delta(\tau-t_0)s(t-\tau)d\tau = s(t-t_0) \tag{2.31}$$

In sum, convolution of $s(t)$ with an impulse function does not change the functional form of $s(t)$. The only thing it may do is cause a time shift in $s(t)$ if the impulse does not occur at $t = 0$.

Now that we have a feel for the operation known as convolution, let us return to our study of the Fourier transform.

The *convolution theorem* states that the Fourier transform of a function of time that is the convolution of two functions of time is equal to the product of the two corresponding Fourier transforms. That is, if

$$r(t) \leftrightarrow R(f)$$
$$s(t) \leftrightarrow S(f)$$

then

$$r(t)*s(t) \leftrightarrow R(f)S(f) \tag{2.32}$$

The proof of the theorem is straightforward. We simply evaluate the Fourier transform of the convolution:

$$\begin{aligned}\mathcal{F}[r(t)*s(t)] &= \int_{-\infty}^{\infty} e^{-j2\pi ft}\left[\int_{-\infty}^{\infty} r(\tau)s(t-\tau)d\tau\right]dt \\ &= \int_{-\infty}^{\infty} r(\tau)\left[\int_{-\infty}^{\infty} e^{-j2\pi ft}s(t-\tau)dt\right]d\tau\end{aligned} \tag{2.33}$$

We now make a change of variables in the inner integral by letting $t - \tau = k$. We then have

$$\mathcal{F}[r(t)*s(t)] = \int_{-\infty}^{\infty} r(\tau)e^{-j2\pi f\tau}\left[\int_{-\infty}^{\infty} s(k)e^{-j2\pi fk}dk\right]d\tau \tag{2.34}$$

The integral in the brackets is simply $S(f)$. Since $S(f)$ is not a function of τ, it can be pulled out of the outer integral. This yields the desired result,

$$\mathcal{F}[r(t)*s(t)] = S(f)\int_{-\infty}^{\infty} r(\tau)e^{-j2\pi f\tau}d\tau = S(f)R(f) \tag{2.35}$$

and the theorem is proved.

Convolution is an operation performed between two functions. These need not be functions of the independent variable t; we could just as easily have convolved two Fourier transforms together to get a third function of f:

$$H(f) = R(f)*S(f) = \int_{-\infty}^{\infty} R(k)S(f-k)dk \tag{2.36}$$

Since the integral defining the Fourier transform and that yielding the inverse transform are quite similar, one might guess that convolution of two transforms corresponds to

multiplication of the two corresponding functions of time. Indeed, one can prove, in an analogous way to the previous proof, that

$$R(f)*S(f) \leftrightarrow r(t)s(t) \tag{2.37}$$

To prove this, simply calculate the inverse Fourier Transform of $R(f)*S(f)$. Equation (2.35) is called the *time convolution theorem* and Eq. (2.37) the *frequency convolution theorem.*

Example 2.11

Use the convolution theorem to evaluate the integral

$$\int_{-\infty}^{\infty} \frac{\sin 3\tau}{\tau} \frac{\sin (t-\tau)}{t-\tau} d\tau$$

Solution: We recognize that the integral represents the convolution

$$\frac{\sin 3t}{t} * \frac{\sin t}{t}$$

The transform of the integral is therefore the product of the transforms of the two functions $(\sin 3t)/t$ and $(\sin t)/t$. These two transforms may be found in Appendix II. They and their product are sketched in Fig. 2.27.

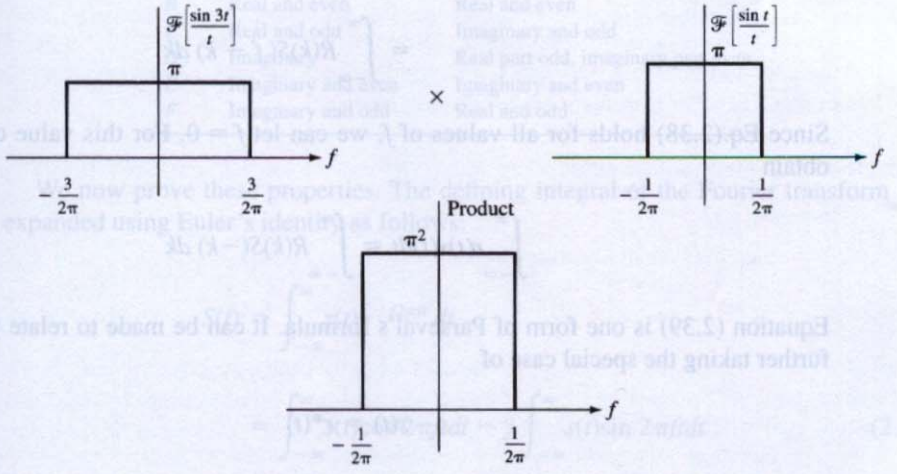


Figure 2.27 Transforms and Product for Example 2.11.

The function of time corresponding to the convolution is simply the inverse transform of the product. This is seen to be

$$\frac{\pi \sin t}{t}$$

Note that when $(\sin t)/t$ is convolved with $(\sin 3t)/t$, the only change that takes place is the addition of a scale factor π. In fact, if $(\sin t)/t$ were convolved with $(\sin 3t)/\pi t$, it would not have changed at all! This surprising result is no accident: There are entire classes of functions that remain unchanged after convolution with $(\sin t)/\pi t$. If this were not true, many of the most basic communication systems could never function.

2.5.2 Parseval's Theorem

There is little similarity between the waveshape of a function and that of its Fourier transform. However, certain relationships do exist between the energy of a function of time and the energy of its transform. Here, we use *energy* to denote the integral of the square of the function. This represents the amount of energy, in watt-seconds, dissipated in a 1Ω resistor if the time signal represents the voltage across or the current through the resistor. Such a relationship proves useful if we know the *transform* of a function of time and wish to know the energy of the function: We do not need to go through the effort of evaluating the inverse transform.

Parseval's theorem, which states this kind of relationship, is derived from the frequency convolution theorem. Starting with that theorem, we have

$$r(t)s(t) \leftrightarrow R(f)*S(f)$$

$$\mathcal{F}[r(t)s(t)] = \int_{-\infty}^{\infty} r(t)s(t)e^{-j2\pi ft}\,dt \tag{2.38}$$

$$= \int_{-\infty}^{\infty} R(k)S(f-k)\,dk$$

Since Eq.(2.38) holds for all values of f, we can let $f = 0$. For this value of f, we then obtain

$$\int_{-\infty}^{\infty} r(t)s(t)dt = \int_{-\infty}^{\infty} R(k)S(-k)\,dk \tag{2.39}$$

Equation (2.39) is one form of Parseval's formula. It can be made to relate to energy by further taking the special case of

$$s(t) = r^*(t)$$

The Fourier transform of the conjugate, $\mathcal{F}[r^*(t)]$, is given by the conjugate of the transform reflected around the vertical axis, that is, $R^*(-f)$. You should take the time now to prove this statement.

Using the preceding result in Eq. (2.39), we find that

$$\int_{-\infty}^{\infty} |r^2(t)|\,dt = \int_{-\infty}^{\infty} |R^2(f)|\,df \tag{2.40}$$

We have used the fact that the product of a function and its complex conjugate is equal to the square of the magnitude of the function.[5] (Convince yourself that the square of the magnitude is the same as the magnitude of the square of a complex number.)

Equation (2.40) shows that the energy of a function of time is equal to the energy of its Fourier transform.

2.6 PROPERTIES OF THE FOURIER TRANSFORM

We now illustrate some of the more important properties of the Fourier transform. One can certainly go through technical life without making use of any of these properties, but to do so would involve considerable repetition and extra work. The properties allow us to derive something *once* and then to use the result for a variety of applications. They also allow us to predict the behavior of various systems.

2.6.1 Real/Imaginary–Even/Odd

The following table summarizes properties of the Fourier transform based upon observations made upon the function of time.

	Function of Time	Fourier Transform
A	Real	Real part even, imaginary part odd
B	Real and even	Real and even
C	Real and odd	Imaginary and odd
D	Imaginary	Real part odd, imaginary part even
E	Imaginary and even	Imaginary and even
F	Imaginary and odd	Real and odd

We now prove these properties. The defining integral of the Fourier transform can be expanded using Euler's identity as follows:

$$S(f) = \int_{-\infty}^{\infty} s(t)e^{-j2\pi ft}\,dt$$

$$= \int_{-\infty}^{\infty} s(t)\cos 2\pi ftdt - j\int_{-\infty}^{\infty} s(t)\sin 2\pi ftdt \tag{2.41}$$

$$= R + jX$$

[5]The time signals we deal with in the real world of communication are real functions of time. However, as in basic circuit analysis, complex mathematical functions are often used to represent sinusoids. A complex number is used for the magnitude and phase angle of a sinusoid. Therefore, although complex signals do not exist in real life, they are often used in "paper" solutions of problems.

R is an even function of f, since, when f is replaced with $-f$, the function does not change. Similarly, X is an odd function of f.

If $s(t)$ is first assumed to be real, R becomes the real part of the transform and X is the imaginary part. Thus, *property A* is proved.

If, in addition to being real, $s(t)$ is even, then $X = 0$. This is true because the integrand in X is odd (the product of an even and an odd function) and integrates to zero. Hence, *property B* is proved.

If $s(t)$ is now real and odd, the same argument applies, but $R = 0$. This proves *property C*.

Now we let $s(t)$ be imaginary. X then becomes the imaginary part of the transform, and R is the real part. From this simple observation, *properties D, E, and F* are easily verified.

2.6.2 Time Shift

The Fourier transform of a shifted function of time is equal to the product of the transform of that function with a complex exponential. That is,

$$s(t - t_0) \leftrightarrow e^{-j2\pi f t_0} S(f) \tag{2.42}$$

Proof. The proof follows directly from evaluation of the transform of $s(t - t_0)$.

$$\mathcal{F}[s(t - t_0)] = \int_{-\infty}^{\infty} s(t - t_0) e^{-j2\pi f t} dt = \int_{-\infty}^{\infty} s(\tau) e^{-j2\pi f(\tau + t_0)} d\tau \tag{2.43}$$

The second integral follows from a change of variables, letting $\tau = (t - t_0)$. We now pull the part that does not depend on τ to the front of the integral and note that the remaining part is a Fourier transform of $s(t)$. Finally, we get

$$\mathcal{F}[s(t - t_0)] = e^{-j2\pi f t_0} S(f) \tag{2.44}$$

Example 2.12

Find the Fourier transform of

$$s(t) = \begin{cases} 1, & 0 < t < 2 \\ 0, & \text{otherwise} \end{cases}$$

The function $s(t)$ is sketched in Fig. 2.28.

Solution: From the definition of the Fourier transform, we have

$$S(f) = \int_0^2 e^{-j2\pi f t} dt = \frac{e^{-j2\pi f}}{j2\pi f} (e^{j2\pi f} - e^{-j2\pi f})$$

$$= e^{-j2\pi f} \frac{\sin 2\pi f}{\pi f}$$

As expected, $S(f)$ is complex, since $s(t)$ is neither even nor odd.

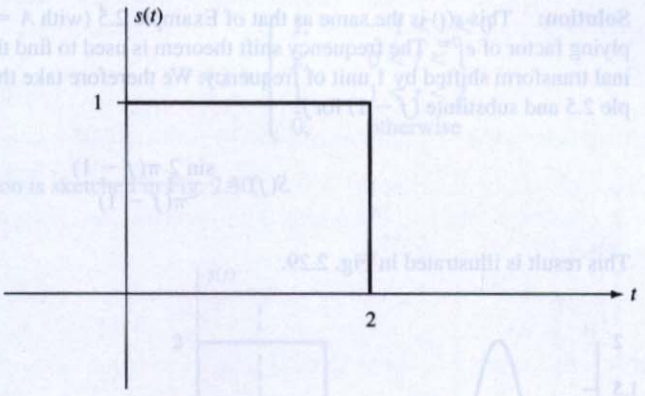


Figure 2.28 $s(t)$ for Example 2.12.

The result of Example 2.12 could have been derived in one step using the answer from Example 2.5 and the time-shift property. The $s(t)$ of Example 2.12 is the same as that of Example 2.5 (with $A = \alpha = 1$) except for a time shift of 1 sec.

2.6.3 Frequency Shift

The function of time corresponding to a shifted Fourier transform is equal to the product of the function of time corresponding to the unshifted transform and a complex exponential. That is,

$$S(f - f_0) \leftrightarrow e^{j2\pi f_0 t} s(t) \tag{2.45}$$

Proof. The proof follows directly from evaluation of the inverse transform of $S(f - f_0)$:

$$\int_{-\infty}^{\infty} S(f - f_0) e^{j2\pi ft} df = \int_{-\infty}^{\infty} S(k) e^{j2\pi t(k + f_0)} dk \tag{2.46}$$

In the second integral, we have made a change of variables, letting $k = f - f_0$. We now pull the part of the integrand that does not depend upon k in front of the integral and recognize that the remaining integral is the inverse Fourier transform of $s(t)$. This yields

$$S(f - f_0) \leftrightarrow e^{j2\pi f_0 t} s(t) \tag{2.47}$$

Example 2.13

Find the Fourier transform of

$$s(t) = \begin{cases} e^{j2\pi t}, & |t| < 1 \\ 0, & \text{otherwise} \end{cases}$$

Solution: This $s(t)$ is the same as that of Example 2.5 (with $A = \alpha = 1$), except for a multiplying factor of $e^{j2\pi t}$. The frequency shift theorem is used to find that the transform is the original transform shifted by 1 unit of frequency. We therefore take the transform found in Example 2.5 and substitute $(f - 1)$ for f:

$$S(f) = \frac{\sin 2\pi(f-1)}{\pi(f-1)}$$

This result is illustrated in Fig. 2.29.

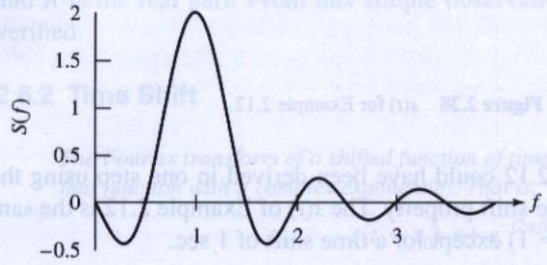


Figure 2.29 $S(f)$ for Example 2.13.

Note that in Example 2.13, the function of time is neither even nor odd. However, the Fourier transform turned out to be a real function of time. Such a situation arises only when the function of time is not real.

2.6.4 Linearity

Linearity is undoubtedly the most important property of the Fourier transform.

The Fourier transform of a linear combination of functions of time is a linear combination of the corresponding Fourier transforms. That is,

$$as_1(t) + bs_2(t) \leftrightarrow aS_1(f) + bS_2(f) \tag{2.48}$$

where a and b are any constants.

Proof. The proof follows directly from the definition of the Fourier transform and from the fact the integration is a linear operation:

$$\int_{-\infty}^{\infty} [as_1(t) + bs_2(t)]e^{-j2\pi ft}dt = a\int_{-\infty}^{\infty} s_1(t)e^{-j2\pi ft}dt + b\int_{-\infty}^{\infty} s_2(t)e^{-j2\pi ft}dt \tag{2.49}$$

$$= aS_1(f) + bS_2(f)$$

Example 2.14

Find the Fourier transform of

$$s(t) = \begin{cases} 1, & -1 < t < 0 \\ 2, & 0 < t < 1 \\ 1, & 1 < t < 2 \\ 0, & \text{otherwise} \end{cases}$$

This function is sketched in Fig. 2.30.

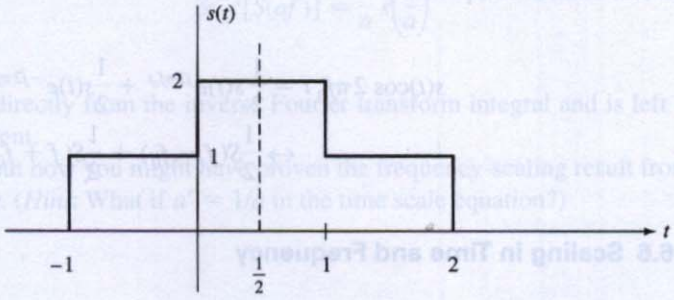


Figure 2.30 $s(t)$ for Example 2.14.

Solution: We use the linearity property and observe that $s(t)$ is the sum of the function of time in Example 2.5 with that in Example 2.12. Therefore, the transform is given by the sum of the two transforms:

$$S(f) = \frac{\sin 2\pi f}{\pi f}[1 + e^{-j2\pi f}]$$

Since the given function of time would be even if shifted to the left by 0.5 sec, we can rewrite this equation in a more descriptive form by factoring out $e^{-j\pi f}$.

$$S(f) = 2\frac{\sin 2\pi f \cos \pi f}{\pi f} e^{-j\pi f}$$

2.6.5 Modulation Theorem

The modulation theorem is very closely related to the frequency shift theorem. We treat it separately because it forms the basis of the entire study of amplitude modulation.

The result of multiplying a function of time by a pure sinusoid is to shift the original transform both up and down by the frequency of the sinusoid (and to cut the amplitude in half).

Proof. We start the proof of this theorem by assuming that $s(t)$ is given, together with its associated Fourier transform. The function $s(t)$ is then multiplied by a cosine waveform to yield

$$s(t)\cos 2\pi f_0 t$$

where the frequency of the cosine is f_0. The Fourier transform of this waveform is given by

$$\mathcal{F}[s(t)\cos 2\pi f_0 t] = \frac{1}{2}S(f - f_0) + \frac{1}{2}S(f + f_0) \tag{2.50}$$

The proof of the modulation theorem follows directly from the frequency shift theorem. We split $\cos 2\pi f_0 t$ into two exponential components and then apply the frequency shift theorem to each component:

$$\begin{aligned} s(t)\cos 2\pi f_0 t &= \frac{1}{2}s(t)e^{j2\pi f_0 t} + \frac{1}{2}s(t)e^{-j2\pi f_0 t} \\ &\leftrightarrow \frac{1}{2}S(f - f_0) + \frac{1}{2}S(f + f_0) \end{aligned} \tag{2.51}$$

2.6.6 Scaling in Time and Frequency

We are rapidly approaching the point of diminishing returns in presenting properties of the Fourier transform. At some point, it is worth dealing with the individual properties as they arise. We shall terminate our exploration with a companion set of two properties referred to as *time and frequency scaling*. The usefulness of these properties arises when you take a function of time or a Fourier transform and either stretch or compress it along the horizontal axis. Thus, if, for example, you already know the Fourier transform of a pulse with width two units (as in Fig. 2.28), you need not do any further calculations to find the Fourier transform of a pulse of any other width.

Time Scaling

Suppose we already know that the Fourier transform of $s(t)$ is $S(f)$. We wish to find the Fourier transform of $s(at)$, where a is a real scaling factor. Thus, if, for example, $a = 2$, we are compressing the function by a factor of 2 along the t-axis, and if $a = 0.5$, we are expanding it by a factor of 2. The result can be derived directly from the definition of the Fourier transform as follows:

$$\mathcal{F}[s(at)] = \int_{-\infty}^{\infty} s(at)e^{-j2\pi ft}\,dt = \int_{-\infty}^{\infty} s(\tau)e^{-j2\pi f\tau/a}\frac{d\tau}{a} \tag{2.52}$$

The latter integral results from a change of variables, letting $s = at$. The integral is now recognized as

$$\mathcal{F}[s(at)] = \frac{1}{a}S\left(\frac{f}{a}\right) \tag{2.53}$$

The result represents a complementary operation on the frequency axis and an amplitude scaling. Thus, if, for example, $a = 2$, the time axis is compressed. In finding the

transform, we expand the frequency axis by a factor of 2 and scale the amplitude of the transform by dividing by 2.

Frequency Scaling

If it is already known that the Fourier transform of $s(t)$ is $S(f)$, then the time signal that has $S(af)$ as its transform, where a is a real scaling factor, is given by

$$\mathcal{F}^{-1}[S(af)] = \frac{1}{a} s\left(\frac{t}{a}\right) \tag{2.54}$$

This is proved directly from the inverse Fourier transform integral and is left as an exercise for the student.

Think about how you might have proven the frequency-scaling result from the time-scaling property. (*Hint*: What if $a' = 1/a$ in the time scale equation?)

2.7 PERIODIC FUNCTIONS

In Example 2.7, the Fourier transform of the cosine function was found to be composed of two impulses, one occurring at the frequency of the cosine and the other at the negative of this frequency. We will now show that the Fourier transform of any periodic function of time is a discrete function of frequency. That is, the transform is nonzero only at discrete points along the f-axis. The proof follows from Fourier series expansions and the linearity of the Fourier transform.

Suppose we find the Fourier transform of a function $s(t)$ that is periodic with period T. We can express the function in terms of the complex Fourier series representation

$$s(t) = \sum_{n=-\infty}^{\infty} c_n e^{jn2\pi f_0 t} \tag{2.55}$$

where

$$f_0 = \frac{1}{T}$$

Previously, we established the transform pair

$$Ae^{j2\pi f_0 t} \leftrightarrow A\delta(f - f_0) \tag{2.56}$$

From this transform pair and the linearity property of the transform, we have

$$\mathcal{F}[s(t)] = \sum_{n=-\infty}^{\infty} c_n \mathcal{F}[e^{jn2\pi f_0 t}] \tag{2.57}$$

This transform is shown in Fig. 2.31 for a representative $s(t)$. Note that the c_n are complex numbers, so the sketch is intended for conceptual purposes only. If the function of time is real and even, the c_n will be real.

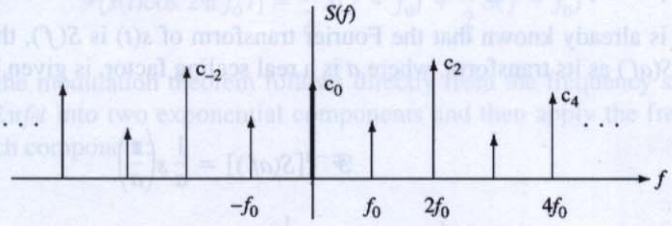


Figure 2.31 Transform of periodic $s(t)$.

The foregoing proof shows that the Fourier transform of a periodic function of time is a train of equally spaced impulses, each of whose strength is equal to the corresponding Fourier coefficient c_n.

Example 2.15

Find the Fourier transform of the periodic function made up of unit impulses, as shown in Fig. 2.32. The function is

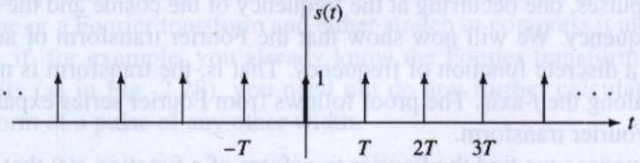


Figure 2.32 Periodic $s(t)$ for Example 2.15.

$$s(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT)$$

Solution: The Fourier transform is found directly from Eq. (2.57). We have

$$S(f) = \sum_{n=-\infty}^{\infty} c_n \delta(f - nf_0)$$

where

$$f_0 = \frac{1}{T}$$

$$c_n = \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} s(t)e^{-jn2\pi f_0 t}dt$$

Within the range of integration, the only contribution of $s(t)$ is that due to the impulse at the origin. Therefore,

$$c_n = \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} \delta(t)e^{-jn2\pi f_0 t}dt = \frac{1}{T}$$

Finally, the Fourier transform of the pulse train is given by

$$S(f) = \frac{1}{T}\sum_{n=-\infty}^{\infty} \delta(f - nf_0)$$

where

$$f_0 = \frac{1}{T}$$

The function of Example 2.15 has an interesting Fourier series expansion. All of the coefficients are equal. Each frequency component possesses the same amplitude as every other component. This is analogous to the observation that the Fourier transform of a single impulse is a constant. The similarity leads us to examine the relationship between the Fourier transform of a periodic function and the Fourier transform of one period of the function.

Suppose that $s(t)$ represents a single period of the periodic function $s_p(t)$. Then we can express the periodic function as a sum of shifted versions of $s(t)$:

$$s_p(t) = \sum_{n=-\infty}^{\infty} s(t - nT) \tag{2.58}$$

Since convolution with an impulse simply shifts the original function, Eq. (2.58) can be rewritten as

$$s_p(t) = s(t) * \sum_{n=-\infty}^{\infty} \delta(t - nT) \tag{2.59}$$

Convolution in the time domain is equivalent to multiplication of the Fourier transforms. The Fourier transform of the train of impulses was found in Example 2.15. Transforming Eq. (2.59) then yields

$$S_p(t) = S(f)\sum_{n=-\infty}^{\infty} \frac{1}{T}\,\delta(f - nf_0) \tag{2.60}$$

Equation (2.60) shows that the Fourier transform of the periodic function is simply a sampled and scaled version of the transform of a single period of the waveform.

PROBLEMS

2.1.1 Evaluate the Fourier series expansion of each of the periodic functions shown in Fig. P2.1.1. Use either the complex exponential or the trigonometric form of the series.

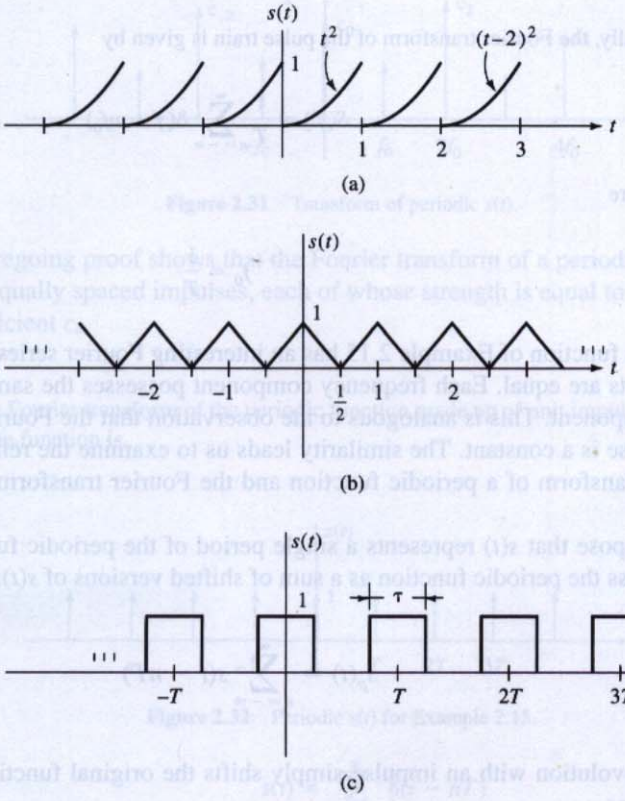


Figure P2.1.1

2.1.2 Evaluate the Fourier series expansion of the function shown in Fig. 2.3.

2.1.3 Evaluate the Fourier series representation of the periodic function

$$s(t) = 2\sin \pi t + 3\sin 2\pi t$$

2.1.4 The periodic function of Fig. P2.1.1(c) is expressed in a trigonometric Fourier series. Find the error if only three terms of the Fourier series are used. Repeat for four terms and five terms.

2.1.5 Find the Fourier series expansion of a gated sinusoid, as shown in Fig. P2.1.5.

2.1.6 The triangular waveform shown in Fig. P2.1.1(b) forms the input to a diode circuit, as shown in Fig. P2.1.6. Assume that the diode is ideal. Find the Fourier series expansion of the current $i(t)$.

2.1.7 In Fig. P2.1.7, find a Fourier series expansion of $s(t)$ that applies for $-\tau/2 < t < \tau/2$.

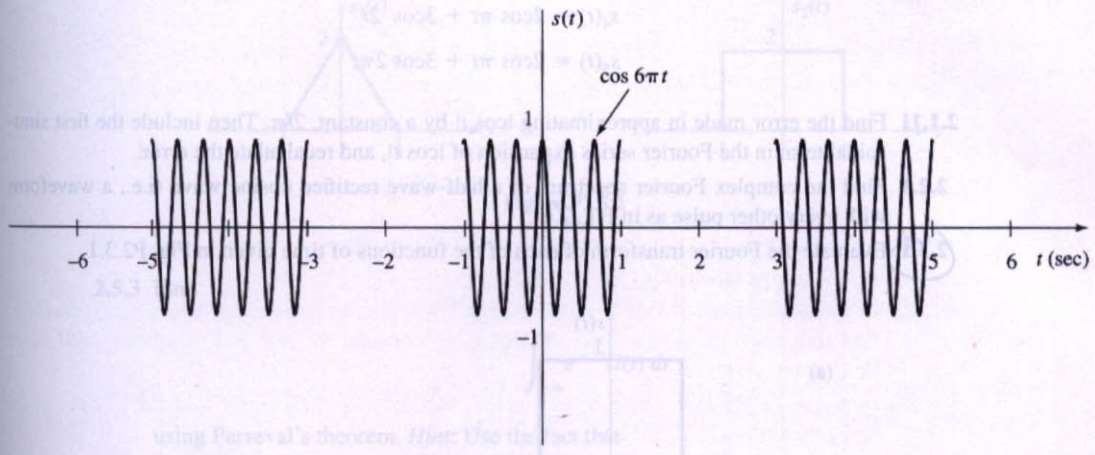


Figure P2.1.5

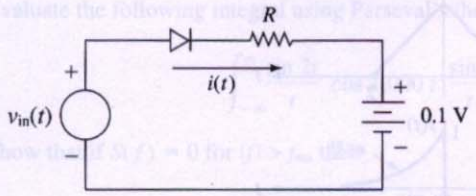


Figure P2.1.6

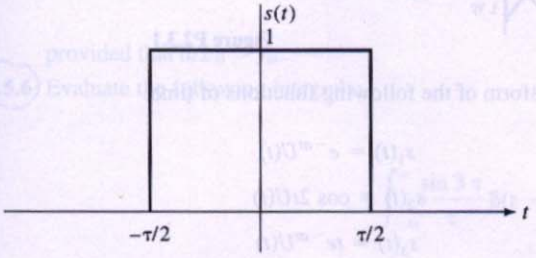


Figure P2.1.7

2.1.8 Find the complex Fourier series representation of $s(t) = t^2$ that applies in the interval $0 < t < 1$. How does this compare with your answer to Problem 2.1.1, Fig. P2.1.1(a)?

2.1.9 Find a trigonometric Fourier series representation of the function

$$s(t) = \cos \pi t$$

in the interval $0 < t < 2$.

2.1.10 Which of the following could *not* be the Fourier series expansion of a periodic signal?

$$s_1(t) = 2\cos t + 3\cos 3t$$

$$s_2(t) = 2\cos 0.5t + 3\cos 3.5t$$

$$s_3(t) = 2\cos 0.25t + 3\cos 0.00054t$$

$$s_4(t) = 2\cos \pi t + 3\cos\ 2t$$

$$s_5(t) = 2\cos \pi t + 3\cos 2\pi t$$

2.1.11 Find the error made in approximating $|\cos t|$ by a constant, $2/\pi$. Then include the first sinusoidal term in the Fourier series expansion of $|\cos t|$, and recalculate the error.

2.2.1 Find the complex Fourier spectrum of a half-wave rectified cosine wave (i.e., a waveform with every other pulse as in Fig. 2.6).

2.3.1 Evaluate the Fourier transform of each of the functions of time given in Fig. P2.3.1.

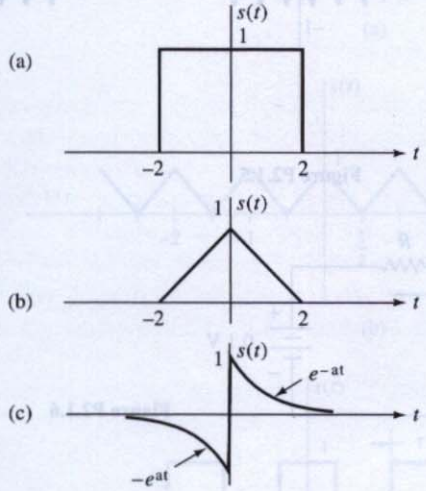


Figure P2.3.1

2.3.2 Evaluate the Fourier transform of the following functions of time:

$$s_1(t) = e^{-at}U(t)$$

$$s_2(t) = \cos 2tU(t)$$

$$s_3(t) = te^{-at}U(t)$$

2.5.1 Convolve $e^{at}U(-t)$ with $e^{-at}U(t)$. These two functions are shown in Fig. P2.5.1.

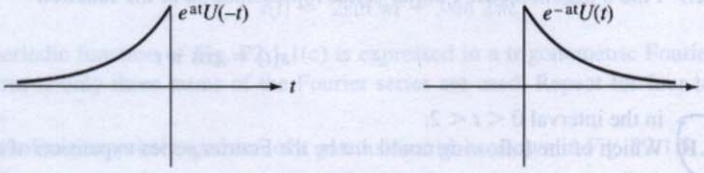


Figure P2.5.1

2.5.2 Convolve together the two functions shown in Fig. P2.5.2, using the convolution integral. Repeat using graphical techniques.

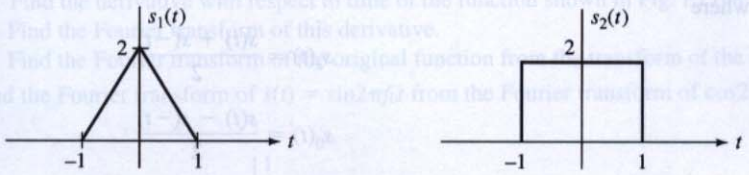


Figure P2.5.2

2.5.3 Find

$$\int_{-\infty}^{\infty} e^{-2t}U(t)\,dt$$

using Parseval's theorem. *Hint*: Use the fact that

$$e^{-2t}U(t) = |e^{-t}U(t)|^2$$

2.5.4 Evaluate the following integral using Parseval's theorem:

$$\int_{-\infty}^{\infty} \frac{\sin 2t}{t} \cos 1{,}000\,t\, \frac{\sin t}{t} \cos 2{,}000\,t\,dt$$

2.5.5 Show that if $S(f) = 0$ for $|f| > f_m$, then

$$s(t) * \frac{\sin at}{\pi t} = s(t)$$

provided that $a/2\pi > f_m$.

2.5.6 Evaluate the following integrals:

$$\int_{-\infty}^{\infty} \frac{\sin 3\tau}{\tau} \delta(t-\tau)\,d\tau$$

$$\int_{-\infty}^{\infty} \frac{\sin 3(\tau-3)}{\tau-3} \frac{\sin 5(t-\tau)}{t-\tau}\,d\tau$$

$$\int_{-\infty}^{\infty} \delta(t-t)(t^3+4)dt$$

2.6.1 **(a)** Write the convolution of $s(t)$ with $U(t)$ in integral form. See whether you can identify this as the integral of $s(t)$.

(b) What is the transform of $s(t)*U(t)$? Solve this using the convolution theorem.

2.6.2 Any arbitrary function can be expressed as the sum of an even and odd function; that is,

$$s(t) = s_e(t) + s_0(t)$$

where

$$s_e(t) = \frac{s(t) + s(-t)}{2}$$

$$s_0(t) = \frac{s(t) - s(-t)}{2}$$

(a) Show that $s_e(t)$ is an even function and that $s_0(t)$ is an odd function.
(b) Show that $s(t) = s_e(t) + s_0(t)$.
(c) Find $s_e(t)$ and $s_0(t)$ for $s(t) = U(t)$, a unit step function.
(d) Find $s_e(t)$ and $s_0(t)$ for $s(t) = \cos 20\pi t$

2.6.3 Given a function $s(t)$ that is zero for negative t, find a relationship between $s_e(t)$ and $s_0(t)$. Can this result be used to find a relationship between the real and imaginary parts of the Fourier transform of $s(t)$?

2.6.4 Evaluate the Fourier transform of $\cos 5\pi t$, starting with the Fourier transform of $\cos \pi t$ and using the time-scaling property.

2.6.5 Given that the Fourier transform of $s(t)$ is $S(f)$:
(a) What is the Fourier transform of ds/dt in terms of $S(f)$?
(b) What is the Fourier transform of

$$\int_{-\infty}^{t} s(\tau)d\tau$$

in terms of $S(f)$?

2.6.6 Use the time shift property to find the Fourier transform of

$$\frac{s(t + T) - s(t)}{T}$$

From this result, find the Fourier transform of ds/dt.

2.6.7 A signal $s(t)$ is put through a gate and truncated in time as shown in Fig. P2.6.7. The gate is closed for $1 < t < 2$. Therefore,

$$s_{out}(t) = \begin{cases} s(t), & 1 < t < 2 \\ 0, & \text{otherwise} \end{cases}$$

Find the Fourier transform of $s_{out}(t)$ in terms of $S(f)$.

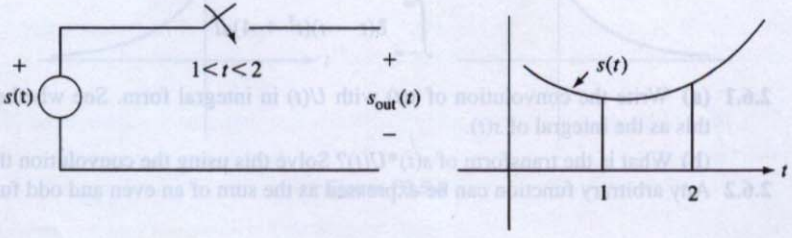


Figure P2.6.7

2.6.8 **(a)** Find the derivative with respect to time of the function shown in Fig. P2.6.8.
(b) Find the Fourier transform of this derivative.
(c) Find the Fourier transform of the original function from the transform of the derivative.

2.6.9 Find the Fourier transform of $s(t) = \sin 2\pi f_0 t$ from the Fourier transform of $\cos 2\pi f_0 t$ and the time shift property.

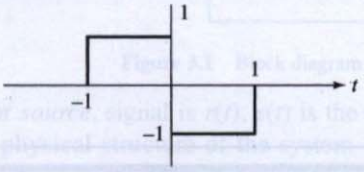


Figure P2.6.8

2.6.10 Find the Fourier transform of $\cos^2 2\pi f_0 t$ from the frequency convolution theorem and the transform of $\cos 2\pi f_0 t$. Check your answer by expanding $\cos^2 2\pi f_0 t$ using trigonometric identities.

2.7.1 The function $s(t)$ of Fig. P2.1.1(c) with $T = 1$ and $\tau = 0.5$ multiplies a function of time, $g(t)$, with $G(f)$ as shown in Fig. P2.7.1(a).
(a) Sketch the Fourier transform of $g_s(t) = g(t)s(t)$.
(b) Can $g(t)$ be recovered from $g_s(t)$?
(c) If $G(f)$ is now as shown in Fig. P2.7.1(b), can $g(t)$ be recovered from $g_s(t)$? Explain your answer.

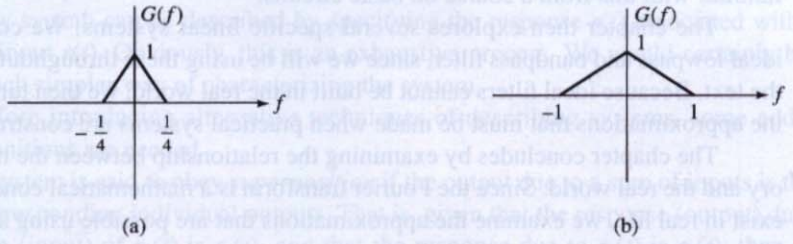


Figure P2.7.1

3

Linear Systems

3.0 PREVIEW

What We Will Cover and Why You Should Care

The previous chapter developed the basic mathematical tools required for waveform analysis. We now apply these techniques to the study of linear systems in order to determine system capabilities and features. We then will be in a position to pick those particular characteristics of linear systems which are desirable for communication applications.

This chapter begins by developing the concepts that are necessary to understand a *system function*, which is a way to describe the behavior of a particular system. We then relate the system function to the sinusoidal steady-state response of circuits. You should be familiar with this from a course on basic circuits.

The chapter then explores several specific linear systems. We concentrate on the ideal lowpass and bandpass filter, since we will be using these throughout the remainder of the text. Because ideal filters cannot be built in the real world, we then turn our attention to the approximations that must be made when practical systems are constructed.

The chapter concludes by examining the relationship between the mathematical theory and the real world. Since the Fourier transform is a mathematical concept that does not exist in real life, we examine the approximations that are possible using a laboratory spectrum analyzer.

Necessary Background

To understand the concepts presented in this chapter, you must be comfortable with the Fourier transform. You may have to review Chapter 2 from time to time.

3.1 THE SYSTEM FUNCTION

We begin by defining some common terms. A *system* is a set of rules that associates an *output* function of time with every *input* function of time. This is shown in block diagram form in Fig. 3.1.

Figure 3.1 Block diagram of system.

The input, or *source*, signal is $r(t)$; $s(t)$ is the output, or *response*, signal due to the input. The actual physical structure of the system determines the exact relationship between $r(t)$ and $s(t)$.

A single-ended arrow is used as a shorthand method of relating an input to its resulting output. That is,

$$r(t) \rightarrow s(t)$$

is read, "an input $r(t)$ causes an output $s(t)$."

For example, suppose the system under study is an electric circuit. Then $r(t)$ could be an input voltage or current signal, and $s(t)$ could be a voltage or current measured anywhere in the circuit. We would not modify the block diagram representation of Fig. 3.1, even though the circuit schematic would have *two* wires for each voltage. The single lines in the figure represent signal flow.

In the special case of a two-terminal electrical network, $r(t)$ could be a sinusoidal input voltage across two terminals, and $s(t)$ could be the current flowing into one of the terminals due to the impressed voltage. In this case, the relationship between $r(t)$ and $s(t)$ is the *complex impedance* between the two terminals of the network.

Any system can be described by specifying the response $s(t)$ associated with *every* possible input $r(t)$. Obviously, this is an exhaustive process. We would certainly hope to find a much simpler way of characterizing the system.

Before introducing alternative techniques of describing systems, some additional basic definitions are needed.

A system is said to obey *superposition* if the output due to a sum of inputs is the sum of the corresponding individual outputs. That is, given that the response (output) due to an excitation (input) of $r_1(t)$ is $s_1(t)$, and that the response due to $r_2(t)$ is $s_2(t)$, then the response due to $r_1(t) + r_2(t)$ is $s_1(t) + s_2(t)$.

Restating this, we may say that a system which obeys superposition has the property that if

$$r_1(t) \rightarrow s_1(t)$$

and

$$r_2(t) \rightarrow s_2(t)$$

then

$$r_1(t) + r_2(t) \rightarrow s_1(t) + s_2(t)$$

Some thought should convince you that in order for a circuit to obey superposition, the source-free, or transient, response (the response due to the initial conditions) must be

zero. (Let $r_2(t) = 0$ to prove this.) In practice, one often replaces a circuit having nonzero initial conditions with one that has zero initial conditions. Additional sources are added to simulate the contributions of the initial conditions.

A concept closely related to superposition is *linearity*. Assume again that $r_1(t) \rightarrow s_1(t)$ and $r_2(t) \rightarrow s_2(t)$. The system is said to be linear if the relationship

$$ar_1(t) + br_2(t) \rightarrow as_1(t) + bs_2(t)$$

holds for all values of the constants a and b. In the remainder of this text, we will use the words *linearity* and *superposition* interchangeably.

A system is said to be *time invariant* if the response due to an input is not dependent upon the actual time of occurrence of the input. That is, a system is time invariant if a time shift in input signal causes an equal time shift in the output waveform. In symbolic form, if

$$r(t) \rightarrow s(t)$$

then

$$r(t - t_0) \rightarrow s(t - t_0)$$

for all real t_0.

A sufficient condition for an electrical network to be time invariant is that its component values do not change with time (assuming unchanging initial conditions). That is, if all resistances, capacitances, and inductances remain constant, then the system is time invariant.

Returning to the task of characterizing a system, we shall see that for a time-invariant linear system, a very simple description is possible. That is, instead of requiring that we know the response due to *every* possible input, it will turn out that we need know only the output for one *test* input.

We showed earlier that convolution of any function with an impulse yields the original function. That is,

$$\begin{aligned} r(t) &= r(t) * \delta(t) \\ &= \int_{-\infty}^{\infty} r(\tau)\delta(t - \tau)d\tau \end{aligned} \tag{3.1}$$

Although one must always use extra caution in working with impulses, let us assume that the integral can be considered a limiting case of a sum, so that

$$r(t) = \lim_{\Delta\tau \to 0} \sum_{n=-\infty}^{\infty} r(n\Delta\tau)\delta(t - n\Delta\tau)\Delta\tau \tag{3.2}$$

Equation (3.2) represents a weighted sum of delayed impulses. Suppose that this weighted sum forms the input to a linear time-invariant system. The output would then be a weighted sum of delayed outputs due to a single impulse.

Suppose now that we know the system's output due to a single impulse. Let us denote that output as $h(t)$, the *impulse response*. Then the output due to the input of Eq. (3.2) is given by

$$s(t) = \lim_{\Delta\tau \to 0} \sum_{n=-\infty}^{\infty} r(n\Delta\tau)h(t - n\Delta\tau)\Delta\tau \tag{3.3}$$

If we take the limit of the sum, the latter becomes an integral, and we have

$$s(t) = \int_{-\infty}^{\infty} r(\tau)h(t - \tau)d\tau = r(t) * h(t) \tag{3.4}$$

Equation (3.4) states that the output due to *any* input is found by convolving that input with the system's response to an impulse. All we need to know about the system is its impulse response. Equation (3.4) is known as the *superposition integral equation.*

The Fourier transform of the impulse is unity. Therefore, in an intuitive sense, the impulse contains all frequencies to an equal degree. This observation hints at the impulse's suitability as a test function for system behavior. On the negative side, it is not possible to produce a perfect impulse in real life. We can only approximate it with a large-amplitude, very narrow pulse.

Taking the Fourier transform of Eq. (3.4) yields

$$\begin{aligned} S(f) &= R(f)H(f) \\ H(f) &= \frac{S(f)}{R(f)} \end{aligned} \tag{3.5}$$

The Fourier transform of the impulse response is thus the ratio of the output Fourier transform to the input Fourier transform. It is given the name *transfer function* or *system function,* and it completely characterizes a linear time-invariant system.

3.2 COMPLEX TRANSFER FUNCTION

Sinusoidal steady-state analysis defines the *complex transfer function* of a system as the ratio of the output phasor to the input phasor. A *phasor* is a complex number representing the amplitude and phase of a sinusoid. The ratio of phasors is a complex function of frequency. In the special case in which the input is a current flowing between two terminals and the output is the voltage across these terminals, the complex transfer function is the *complex impedance* between the two terminals.

As an example, consider the circuit of Fig. 3.2, where $i_1(t)$ is the input and $v(t)$ is the output. The transfer function is given by

$$H(f) = \frac{4j\pi f}{1 + 4j\pi f} \tag{3.6}$$

Alternatively, if $i_2(t)$ is the output and $i_1(t)$ is the input, the transfer function becomes

$$H(f) = \frac{1}{1 + 4j\pi f} \tag{3.7}$$

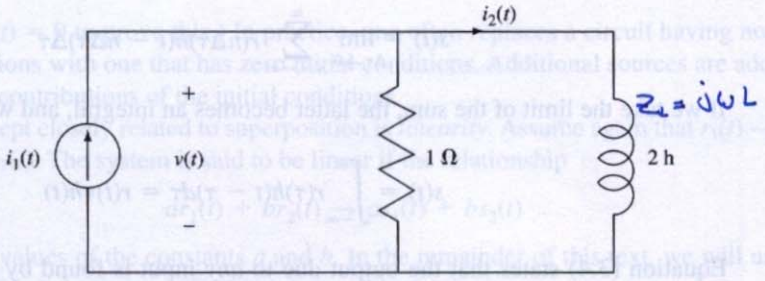


Figure 3.2 Circuit to illustrate transfer function.

We have used the same symbol, $H(f)$, for the transfer function as was used in the previous section to denote the Fourier transform of the impulse response. This is not accidental: *The two expressions are identical.* This statement can be proven by considering a system input of

$$r(t) = e^{j2\pi f_0 t} \tag{3.8}$$

This input is not physically realizable, because it is a complex function of time. Nonetheless, the system equations apply to complex inputs. We can compare the system output using sinusoidal steady-state analysis with that obtained using Fourier transform analysis. Note that the Fourier transform of the complex input function is a shifted impulse. In this manner, we can show that the two $H(f)$ functions are identical.

Example 3.1

In the circuit of Fig. 3.3, the capacitor is initially uncharged. Find $i(t)$, assuming that $v(t) = \delta(t)$.

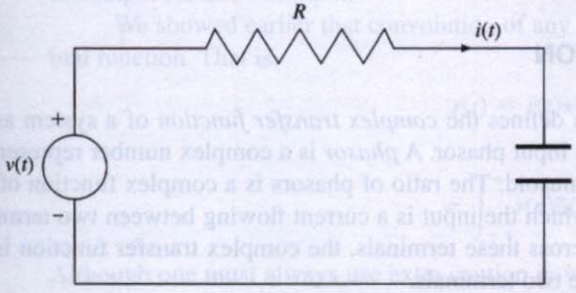


Figure 3.3 Circuit for Example 3.1.

Solution: Since the input to the circuit is an impulse, we are being asked to find the impulse response, $h(t)$.

$H(f)$ can be found using sinusoidal steady-state analysis, where the capacitor is replaced by an impedance of $1/j2\pi fC$. The transfer function is then simply the reciprocal of the circuit input impedance:

$$H(f) = \frac{1}{R + 1/j2\pi fC} = \frac{j2\pi fC}{1 + j2\pi fRC}$$

We now need to find the inverse Fourier transform of $H(f)$. There are a number of ways to do this. One way is to look in a table of Fourier transforms, such as the one in Appendix II of this text. Indeed, there are much more extensive tables available in various handbooks, and you might be fortunate enough to find an entry that applies to your problem. A second technique is to try to evaluate the inverse Fourier transform integral, either in closed form or using computer approximations (e.g., *Mathcad*). Yet another way is to use the FFT. To do so, however, you must become aware of its properties and the relationship between frequency sampling and time sampling.

We shall use tables to solve this particular problem. We rewrite the equations for $H(f)$ borrowing the expansion technique commonly used in Laplace transform analysis:

$$H(f) = \frac{1}{R} - \frac{1/R}{1 + j2\pi fRC}$$

$$h(t) = i(t) = \frac{1}{R}\delta(t) - \frac{1}{R^2C}e^{-t/RC}U(t)$$

This waveform is illustrated in Fig. 3.4.

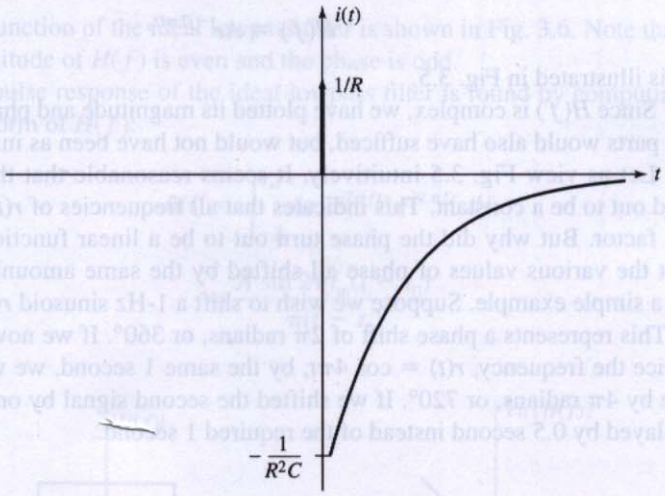


Figure 3.4 $i(t)$ for Example 3.1.

Note that the impulse in $i(t)$ has appeared without additional analysis effort. This is noteworthy, since classical circuit analysis techniques handle impulses with a great deal of difficulty.

The result of Example 3.1 is valid only for zero initial charge on the capacitor. Otherwise, superposition is violated (prove it!), and the output is not the convolution of $h(t)$ with the input.

This apparent shortcoming of Fourier transform analysis of systems with nonzero initial conditions is circumvented by treating initial conditions as sources. The consideration of initial conditions is not critical to most communication systems, so we will usually assume zero initial conditions.

3.3 FILTERS

In ordinary language, the word *filter* refers to the removal of the undesired parts of something. In linear system theory, it was probably originally applied to systems that eliminate undesired frequency components from a time waveform. The term has evolved to include systems that simply weight the various frequency components of a signal.

Many of the communication systems we discuss contain *ideal distortionless filters*. We therefore begin our study by defining *distortion*.

A *distorted* time signal is a time signal whose basic shape has been altered. $r(t)$ can be multiplied by a constant and shifted in time without changing the basic shape of the waveform.

In mathematical terms, we consider $Ar(t - t_0)$, where A and t_0 are any real constants, to be an undistorted version of $r(t)$. Of course, A cannot equal zero. The Fourier transform of $Ar(t - t_0)$ is found from the time shift property:

$$Ar(t - t_0) \leftrightarrow Ae^{-j2\pi f t_0}R(f) \tag{3.9}$$

We can consider this the output of a linear system with input $r(t)$ and system function

$$H(f) = Ae^{-j2\pi f t_0} \tag{3.10}$$

This is illustrated in Fig. 3.5.

Since $H(f)$ is complex, we have plotted its magnitude and phase. The real and imaginary parts would also have sufficed, but would not have been as instructive.

Let us view Fig. 3.5 intuitively. It seems reasonable that the magnitude function turned out to be a constant. This indicates that all frequencies of $r(t)$ are multiplied by the same factor. But why did the phase turn out to be a linear function of frequency? Why aren't the various values of phase all shifted by the same amount? The answer is clear from a simple example. Suppose we wish to shift a 1-Hz sinusoid $r(t) = \cos 2\pi t$ by 1 second. This represents a phase shift of 2π radians, or 360°. If we now wish to shift a signal of twice the frequency, $r(t) = \cos 4\pi t$, by the same 1 second, we would have to shift the phase by 4π radians, or 720°. If we shifted the second signal by only 360°, it would only be delayed by 0.5 second instead of the required 1 second.

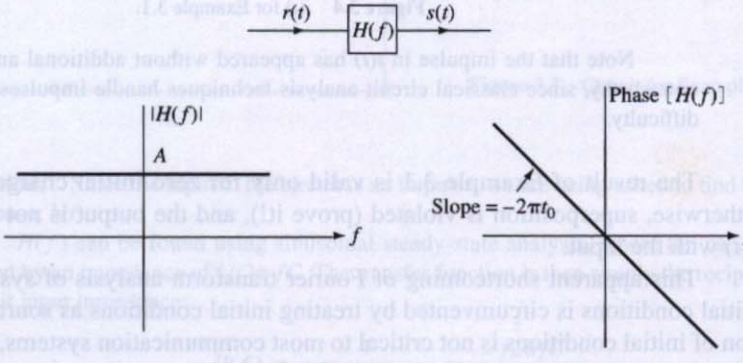


Figure 3.5 Characteristics of distortionless system.

Now consider a general signal composed of many frequency components. If we delay all components by the same angular phase, we would not be delaying them by the same amount of time, and the signal would be severely distorted. In order to delay by the same amount of time, the phase shift must be proportional to frequency.

3.3.1 Ideal Lowpass Filter

An *ideal lowpass filter* is a linear system that acts like an ideal distortionless system, provided that the input signal contains no frequency components above the *cutoff* frequency of the filter. Frequency components above this cutoff are completely blocked from appearing at the output. The cutoff frequency is the maximum frequency passed by the filter, and we denote it as f_m. The system function is then given by

$$H(f) = \begin{cases} Ae^{-j2\pi f t_0}, & |f| < f_m \\ 0, & |f| > f_m \end{cases} \tag{3.11}$$

The transfer function of the ideal lowpass filter is shown in Fig. 3.6. Note that since $h(t)$ is real, the magnitude of $H(f)$ is even and the phase is odd.

The impulse response of the ideal lowpass filter is found by computing the inverse Fourier transform of $H(f)$:

$$\begin{aligned} h(t) &= \int_{-f_m}^{f_m} Ae^{-j2\pi f t_0} e^{j2\pi f t}\, df \\ &= \frac{A \sin 2\pi f_m (t - t_0)}{\pi (t - t_0)} \end{aligned} \tag{3.12}$$

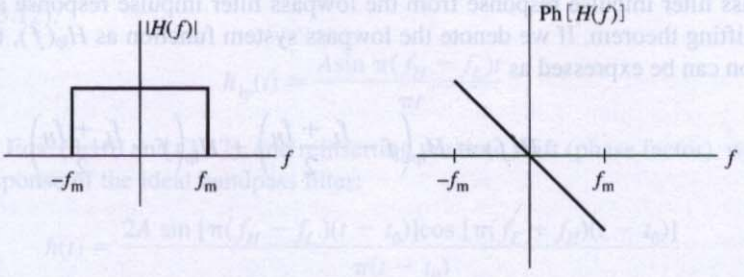


Figure 3.6 Ideal lowpass filter characteristic.

This impulse response is shown in Fig. 3.7. The amount of delay, t_0, is proportional to the slope of the phase characteristic. The cutoff frequency is proportional to the peak of $h(t)$ and inversely proportional to the spacing between zero-axis crossings of the function. That is, as f_m increases, the peak of $h(t)$ increases, and the width of the shaped pulse decreases—the response gets taller and skinnier.

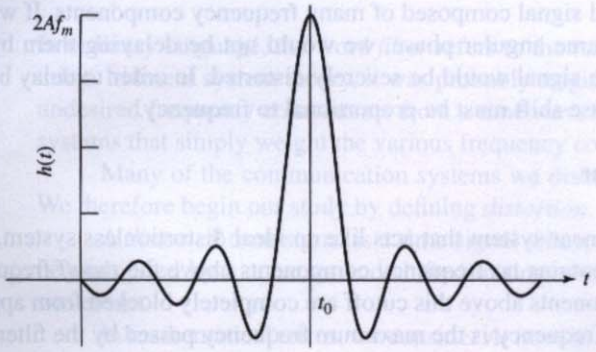


Figure 3.7 Impulse response of ideal lowpass filter.

3.3.2 Ideal Bandpass Filter

Rather than pass frequencies between zero and f_m, as in the case of the lowpass filter, the ideal bandpass filter passes frequencies between two nonzero frequencies, f_L and f_H. The filter acts like an ideal distortionless system, provided that the input signal contains no frequency components outside of the filter *passband.* The system function of the ideal bandpass filter is

$$H(f) = \begin{cases} Ae^{-j2\pi ft_0}, & f_L < |f| < f_H \\ 0, & \text{otherwise} \end{cases} \tag{3.13}$$

This function is illustrated in Fig. 3.8.

The impulse response of the bandpass filter can be found by evaluating the inverse Fourier transform of $H(f)$. Alternatively, we can save a lot of work by deriving the bandpass filter impulse response from the lowpass filter impulse response and the frequency-shifting theorem. If we denote the lowpass system function as $H_{lp}(f)$, the bandpass function can be expressed as

$$H(f) = H_{lp}\left(f - \frac{f_L + f_H}{2}\right) + H_{lp}\left(f + \frac{f_L + f_H}{2}\right) \tag{3.14}$$

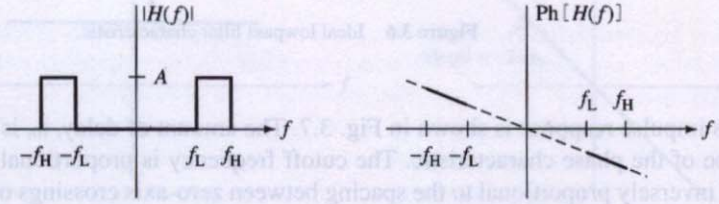


Figure 3.8 Ideal bandpass filter characteristic.

Figure 3.9 shows the relationship between $H(f)$ and $H_{lp}(f)$.

We have illustrated the system functions as if they were real functions of frequency. That is, for purposes of the derivation, we are assuming that $t_0 = 0$. This approach is justified by the time invariance of the system. When we are finished, we can simply insert a time shift and the associated phase factor. Alternatively, you can view Fig. 3.9 as a plot of the magnitudes of the functions and carry the exponential phase term through every step of the derivation.

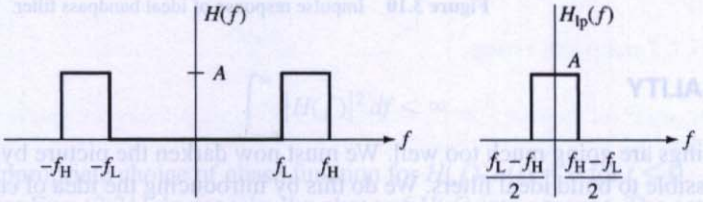


Figure 3.9 Bandpass and lowpass characteristics.

If we define the midpoint of the passband (the average of f_L and f_H) as

$$f_{av} = \frac{f_L + f_H}{2} \tag{3.15}$$

then the impulse response is

$$\begin{aligned} h(t) &= h_{lp}(t)e^{j2\pi f_{av}t} + h_{lp}(t)e^{-j2\pi f_{av}t} \\ &= 2h_{lp}(t)\cos 2\pi f_{av}t = 2h_{lp}(t)\cos\left[\pi(f_L + f_H)t\right] \end{aligned} \tag{3.16}$$

From Eq. (3.12),

$$h_{lp}(t) = \frac{A\sin \pi(f_H - f_L)t}{\pi t} \tag{3.17}$$

Combining Eqs. (3.16) and (3.17), and reinserting the time shift (phase factor), we find the impulse response of the ideal bandpass filter:

$$h(t) = \frac{2A\sin\left[\pi(f_H - f_L)(t - t_0)\right]\cos\left[\pi(f_L + f_H)(t - t_0)\right]}{\pi(t - t_0)} \tag{3.18}$$

The impulse response is illustrated in Fig. 3.10. The outline of this waveform resembles the impulse response of the lowpass filter. Note that as the two limiting frequencies become large compared to the difference between them, the impulse response starts resembling a shaded-in version of the lowpass impulse response and its mirror image. This happens when the center frequency of the bandpass filter becomes large compared to the width of its passband. This observation will prove significant in our later studies of amplitude modulation.

Figure 3.10 Impulse response of ideal bandpass filter.

3.4 CAUSALITY

Things are going much too well. We must now darken the picture by showing that it is impossible to build ideal filters. We do this by introducing the idea of *causality*. This refers to the cause-and-effect relationship. The effect, or response, due to a cause, or input, cannot anticipate the input. That is, a causal system's output at any particular time depends only upon the input prior to that time, and not upon any future values of the input. There are no crystal balls in the real technical world.[1] For a linear system to be causal, it is necessary and sufficient that the impulse response $h(t)$ be zero for $t < 0$. It follows that the response due to a general input $r(t)$ depends only on past values of $r(t)$. To verify this, we write the expression for the output of a linear system in terms of its input and impulse response:

$$s(t) = \int_{-\infty}^{\infty} h(\tau)r(t-\tau)d\tau \tag{3.19}$$

If $h(t) = 0$ for $t < 0$, Eq. (3.19) becomes

$$s(t) = \int_{0}^{\infty} h(\tau)r(t-\tau)d\tau = \int_{-\infty}^{t} r(k)h(t-k)dk \tag{3.20}$$

The second integral in this equation results from a change of variables, where we let $k = t - \tau$. Equation (3.20) clearly shows that the output depends only on *past* values of the input. This proves sufficiency. To prove necessity, we note that $\delta(t) = 0$ for all $t < 0$. In a causal system, the inputs $r(t) = 0$ and $r(t) = \delta(t)$ must yield the same outputs, at least until time $t = 0$. That is, if the system cannot "anticipate" future values of the input, it has no way of "telling" the difference between zero and $\delta(t)$ prior to time $t = 0$. But in a linear system, an input that is identically equal to zero yields an output that is also identically equal to zero. Therefore, $h(t)$ must equal zero for $t < 0$, and the necessity part of the statement is proven.

The study of causality is important because, in general, causal systems are physically realizable and noncausal systems are not (much to the dismay of astrologers, fortune-tellers, and lottery players).

[1]There is a class of systems known as *prediction filters*. In fact, we use these later in the text in discussing both source encoders for digital systems and data compression of speech signals. Although we use the word *prediction*, there is nothing mysterious or noncausal about such systems.

The foregoing criterion is easy to apply if $h(t)$ is explicitly known. That is, one need simply examine $h(t)$ to see if it is zero for negative t. In many cases, however, it will be $H(f)$ that is known, and the Fourier inversion to find $h(t)$ may be difficult to perform. It would therefore be helpful if the constraint on $h(t)$ could be translated into a constraint on $H(f)$. We could then determine whether a system could be built without having to invert $H(f)$.

The *Paley-Wiener criterion* states that if

$$\int_{-\infty}^{\infty} \frac{|\ln H(f)|}{1 + (2\pi f)^2}\, df < \infty \tag{3.21a}$$

and

$$\int_{-\infty}^{\infty} |H(f)|^2\, df < \infty \tag{3.21b}$$

then, for an appropriate choice of phase function for $H(f)$, $h(t) = 0$ for $t < 0$.

Note that Eqs. (3.21) do not take the phase of $H(f)$ into account. The actual form of $h(t)$ certainly does depend upon the phase of $H(f)$. Indeed, if a particular $H(f)$ corresponds to a causal system, we can make that system noncausal by shifting the original impulse response to the left on the time axis. This shift corresponds to a linear change in the phase of $H(f)$, which does not affect Eqs. (3.21). Thus, the Paley-Wiener criterion really tells us whether it is *possible* for $H(f)$ to be the transform of a causal time function (i.e., it is a necessary, but not sufficient, condition). Assuming that Eqs. (3.21) were satisfied, the phase of $H(f)$ would still have to be examined before final determination about the causality of the system could be made. Because of these considerations, we will use the Paley-Wiener condition to make only one simple, but significant, observation.

Since the logarithm of zero approaches minus infinity, if $H(f) = 0$ for any nonzero interval along the f-axis, then the first integral in Eq. (3.21a) does not converge. Therefore, the system cannot be causal. Alas, all of our ideal lowpass and bandpass filters are noncausal. Of course, we already knew this from observing the impulse responses $h(t)$. For this reason, practical filters must be, at best, approximations to their ideal counterparts.

The previous observation about $H(f)$ is a special case of a much more general theorem. Suppose $h(t)$ is such that

$$\int_{-\infty}^{\infty} |h(t)|^2\, dr < \infty \tag{3.22}$$

Then if $h(t)$ is identically zero for any finite range along the t-axis, its Fourier transform cannot be zero for any finite range along the f-axis, and vice versa. This is sometimes stated as "a function that is time limited cannot be bandlimited."

The foregoing statement leads to an important observation: Any function of time that does not exist for all time (i.e., that is zero for any time interval) cannot be bandlimited.

In the next section, we discuss some of the most common approximations that are used in place of ideal filters. We will show that, provided that enough electrical elements are available, the ideal characteristics can be approached arbitrarily closely.

3.5 PRACTICAL FILTERS

We now present circuits that approximate the ideal lowpass and bandpass filters. Throughout this section, we assume that the closer $H(f)$ approaches the system function of an ideal filter, the more the filter will behave in an ideal manner in our applications. This fact is not at all obvious. A small change in $H(f)$ can lead to relatively large changes in $h(t)$. One can examine the consequences and categorize the effects of deviation from the constant-amplitude characteristic or from the linear phase characteristic of the ideal system function. (See the discussion of distortion in Section 1.2.1.)

We begin by analyzing the lowpass filter.

3.5.1 Lowpass Filter

The simplest passive approximation to a lowpass filter is the single energy-storage device circuit. An example is the RC circuit of Fig. 3.11. If the output is taken across the capacitor, this circuit approximates a lowpass filter. The reason is that, as the frequency increases, the capacitor behaves as a short circuit. The transfer function is

$$H(f) = \frac{1/j2\pi fC}{R + 1/j2\pi fC} = \frac{1}{1 + j2\pi fRC} \tag{3.23}$$

The magnitude and phase are

$$\begin{aligned} |H(f)| &= \frac{1}{\sqrt{1 + (2\pi fRC)^2}} \\ \theta(f) &= -\tan^{-1}(2\pi fRC) \end{aligned} \tag{3.24}$$

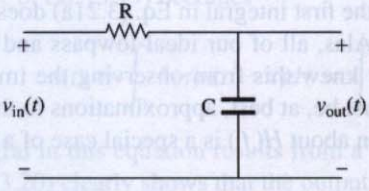
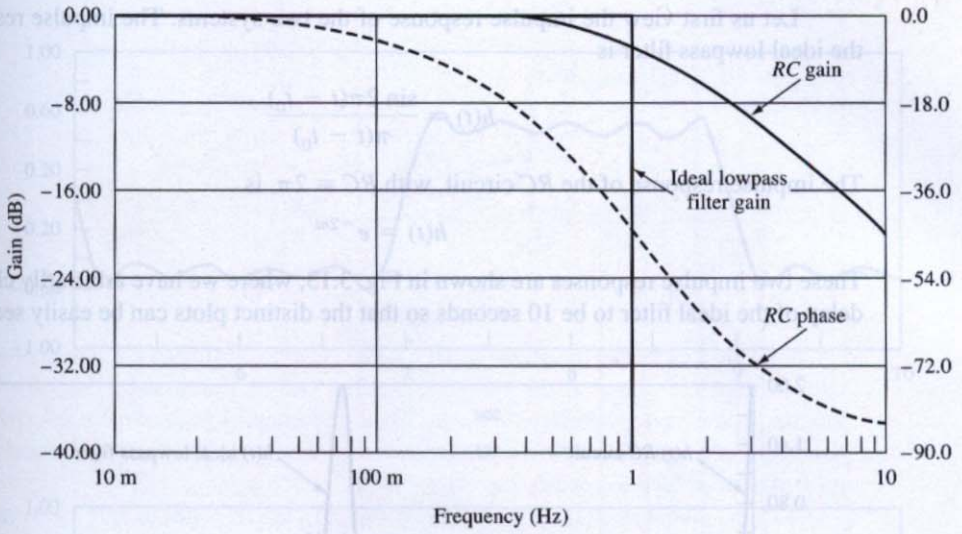


Figure 3.11 *RC* circuit lowpass filter.

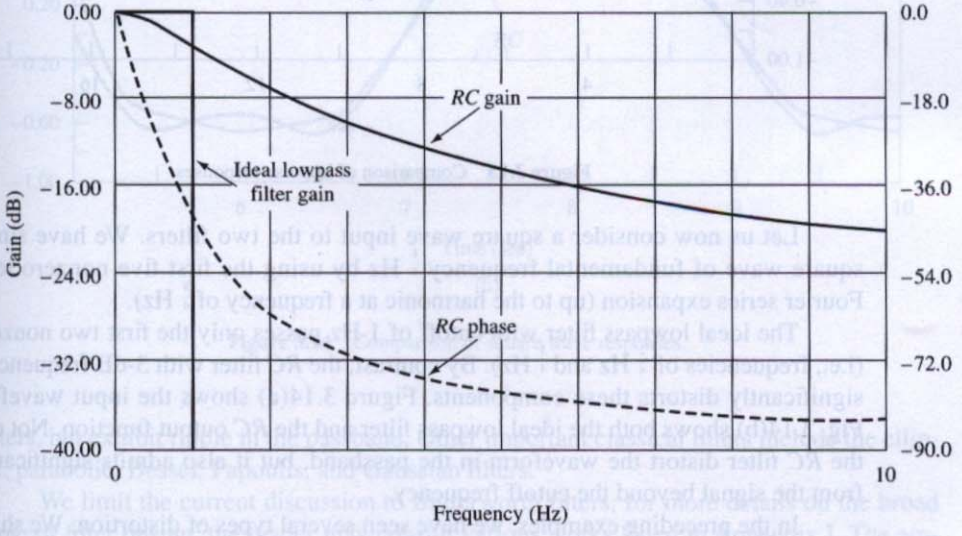
If we set RC to $1/2\pi$, the magnitude of the transfer function drops to $1/\sqrt{2}$ at a frequency of 1 Hz. This is the 3-dB cutoff frequency of the filter ($20 \log(1/\sqrt{2})$, which is approximately -3 dB).[2]

Figure 3.12 shows the magnitude and phase of the RC circuit transfer function. In Fig. 3.12(a), we use a logarithmic frequency axis, while in Fig. 3.12(b) we use a linear frequency axis. Superimposed on each set of curves is the equivalent gain curve for an ideal lowpass filter with a cutoff frequency of 1 Hz. In particular, if we view the linear frequency plot, we see that there is a dramatic difference between the RC approximation and the ideal lowpass filter characteristic. (Keep in mind that in a logarithmic gain curve, a drop of 20 dB is a decrease by a factor of 10.)

[2]The *decibel* (*dB*) is 20 times the logarithm of the amplitude ratio.



(a)



(b)

Figure 3.12 Characteristic of *RC* circuit.

These curves, and the following response waveforms, were developed using MICRO-CAP IV, a computer simulation program. Similar curves would result using other SPICE-based computer simulation programs or if we skip the simulation and plot the functions in Eq. (3.23) using *Mathcad* or MATLAB.

We could continue to analyze the *RC* filter distortion using the techniques derived earlier in this chapter. We choose instead to contrast the output of the *RC* circuit with that of an ideal lowpass filter for several representative inputs.

Let us first view the impulse response of the two systems. The impulse response of the ideal lowpass filter is

$$h(t) = \frac{\sin 2\pi(t - t_0)}{\pi(t - t_0)} \tag{3.25}$$

The impulse response of the *RC* circuit, with $RC = 2\pi$, is

$$h(t) = e^{-2\pi t} \tag{3.26}$$

These two impulse responses are shown in Fig. 3.13, where we have arbitrarily chosen the delay of the ideal filter to be 10 seconds so that the distinct plots can be easily seen.

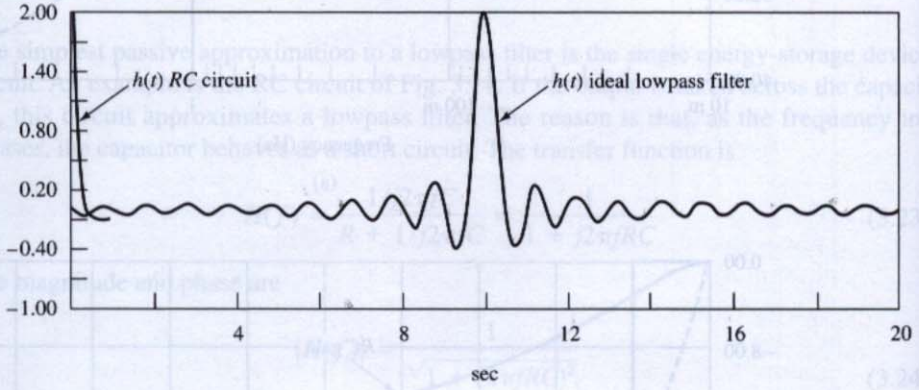


Figure 3.13 Comparison of impulse responses.

Let us now consider a square wave input to the two filters. We have simulated a square wave of fundamental frequency $\frac{1}{4}$ Hz by using the first five nonzero terms in a Fourier series expansion (up to the harmonic at a frequency of $\frac{9}{4}$ Hz).

The ideal lowpass filter with cutoff of 1 Hz passes only the first two nonzero terms (i.e., frequencies of $\frac{1}{4}$ Hz and $\frac{3}{4}$ Hz). By contrast, the *RC* filter with 3-dB frequency at 1 Hz significantly distorts these components. Figure 3.14(a) shows the input waveform, and Fig. 3.14(b) shows both the ideal lowpass filter and the *RC* output function. Not only does the *RC* filter distort the waveform in the passband, but it also admits significant energy from the signal beyond the cutoff frequency.

In the preceding examples, we have seen several types of distortion. We shall revisit distortion later, in the context of amplitude modulation, where we shall use the concepts of group and phase delay to gain an intuitive feel for the effects of the channel on a transmitted waveform. At this point, we would probably agree that an *RC* network is not a very good lowpass filter, except in limited applications. This leads us to explore more complex forms of practical filters.

There are several types of approximations to the ideal lowpass filter, each exhibiting unique characteristics. *Butterworth filters* produce no *ripple* in the passband and attenuate unwanted frequencies outside of this band. They are known as *maximally flat* filters, since they are designed to force the maximum number of derivatives of $H(f)$ (at $f = 0$) to be zero. *Chebyshev filters* attenuate unwanted frequencies more effectively than Butterworth

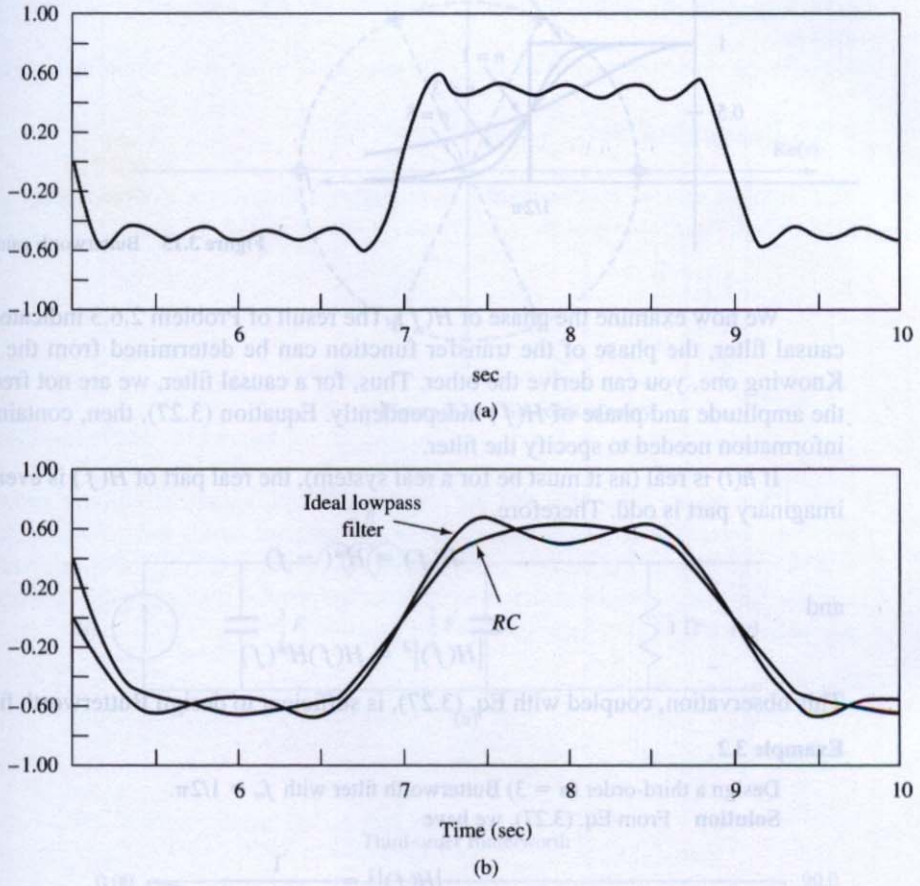


Figure 3.14 Comparison of square wave responses.

filters, but exhibit ripple in the passband. Other important classical filters include the elliptic, parabolic, Bessel, Papoulis, and Gaussian filters.

We limit the current discussion to Butterworth filters; for more details on the broad topic of filter design, the reader may consult various works listed in Appendix I. The amplitude characteristic of the ideal lowpass filter can be approximated by the function

$$|H_n(f)| = \frac{1}{\sqrt{1 + (2\pi f)^{2n}}} \tag{3.27}$$

This function is sketched, for several values of n, in Fig. 3.15. We have illustrated only the positive half of the f-axis, since the function is even. We have chosen $f_m = 1/2\pi$ (1 radian/sec) for the illustration, but a simple scaling process can be used to design a filter for any cutoff frequency. Note that as n gets larger, the amplitude characteristic approaches that of the ideal lowpass filter.

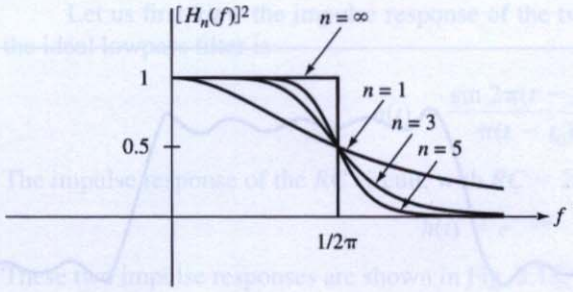


Figure 3.15 Butterworth gain functions.

We now examine the phase of $H(f)$. The result of Problem 2.6.3 indicates that, for a causal filter, the phase of the transfer function can be determined from the amplitude. Knowing one, you can derive the other. Thus, for a causal filter, we are not free to choose the amplitude and phase of $H(f)$ independently. Equation (3.27), then, contains all of the information needed to specify the filter.

If $h(t)$ is real (as it must be for a real system), the real part of $H(f)$ is even, while the imaginary part is odd. Therefore,

$$H(f) = H^*(-f) \tag{3.28a}$$

and

$$|H(f)|^2 = H(f)H^*(f) \tag{3.28b}$$

This observation, coupled with Eq. (3.27), is sufficient to design Butterworth filters.

Example 3.2

Design a third-order ($n = 3$) Butterworth filter with $f_m = 1/2\pi$.

Solution From Eq. (3.27), we have

$$|H(f)|^2 = \frac{1}{1 + (2\pi f)^6}$$

Suppose we change this to the Laplace transform by letting $s = j2\pi f$. We will then be in a position to make observations relative to the poles and zeros of the function. We have

$$|H(s)|^2 = H(s)H(-s) = \frac{1}{1 - s^6}$$

The poles of $|H(s)|^2$ are the six roots of unity, as sketched in Fig. 3.16. They are equally spaced around the unit circle. Three of the poles are associated with $H(s)$ and the other three with $H(-s)$. Since the filter is causal, we associate the three poles in the left half-plane with $H(s)$. $H(s)$ is then found from its poles to be

$$H(s) = \frac{1}{(s - p_1)(s - p_2)(s - p_3)} = \frac{1}{s^3 + 2s^2 + 2s + 1}$$

Given $H(s)$, there are well-known techniques for synthesizing a circuit. If $v(t)$ is the response and $i(t)$ the source, the preceding system function corresponds to the circuit of Fig. 3.17(a). The component values are not realistic, since the cutoff frequency is very low. Figure 3.17(b) shows the computer-simulated frequency characteristic of this circuit.

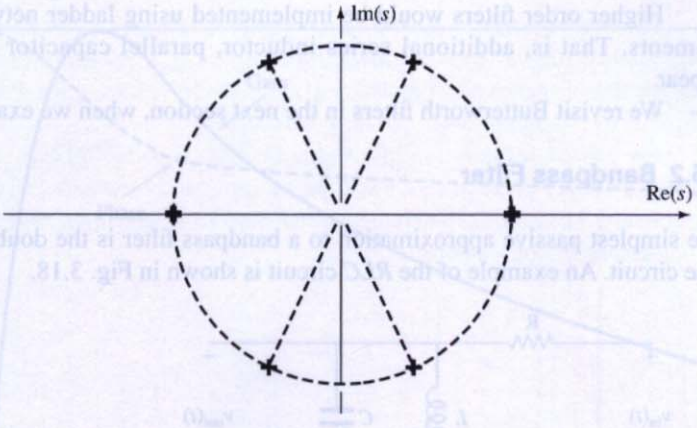
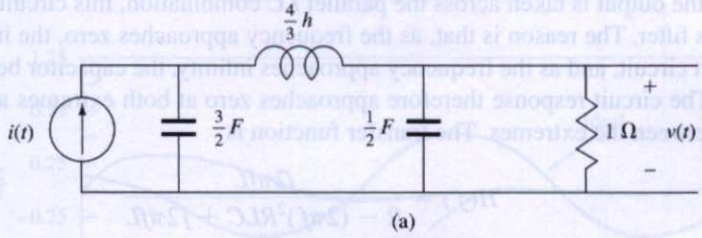
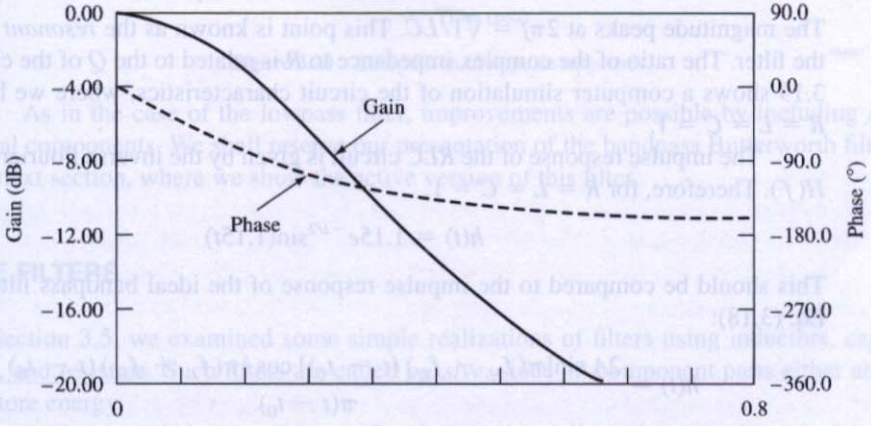


Figure 3.16 Six roots of unity.



(a)

Third-order Butterworth



(b)

Figure 3.17 Third-order Butterworth filter.

Higher order filters would be implemented using ladder networks with additional elements. That is, additional series inductor, parallel capacitor combinations would appear.

We revisit Butterworth filters in the next section, when we examine active filters.

3.5.2 Bandpass Filter

The simplest passive approximation to a bandpass filter is the double energy-storage device circuit. An example of the *RLC* circuit is shown in Fig. 3.18.

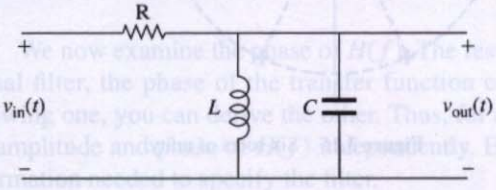


Figure 3.18 *RLC* bandpass circuit.

If the output is taken across the parallel *LC* combination, this circuit approximates a bandpass filter. The reason is that, as the frequency approaches zero, the inductor behaves as a short circuit, and as the frequency approaches infinity, the capacitor behaves as a short circuit. The circuit response therefore approaches zero at both extremes and peaks somewhere between the extremes. The transfer function is

$$H(f) = \frac{j2\pi fL}{R - (2\pi f)^2 RLC + j2\pi fL} \tag{3.29}$$

The magnitude of this is

$$|H(f)| = \frac{1}{\sqrt{R^2[1/2\pi fL) - 2\pi fC]^2 + 1}} \tag{3.30}$$

The magnitude peaks at $2\pi f = \sqrt{1/LC}$. This point is known as the *resonant frequency* of the filter. The ratio of the complex impedance to R is related to the Q of the circuit. Figure 3.19 shows a computer simulation of the circuit characteristics, where we have selected $R = L = C = 1$.

The impulse response of the *RLC* circuit is given by the inverse Fourier transform of $H(f)$. Therefore, for $R = L = C = 1$,

$$h(t) = 1.15e^{-t/2}\sin(1.15t) \tag{3.31}$$

This should be compared to the impulse response of the ideal bandpass filter derived in Eq. (3.18):

$$h(t) = \frac{2A\sin[\pi(f_H - f_L)(t - t_0)]\cos[\pi(f_L + f_H)(t - t_0)]}{\pi(t - t_0)}$$

Figure 3.20 shows the impulse response of the *RLC* circuit and the impulse response of the ideal bandpass filter, where we have chosen $f_H = 0.1$ Hz and $f_L = 0.25$ Hz, the 3-dB points of the *RLC* circuit response. Note that the Q of this filter is extremely low, since the ratio of the bandwidth to the center frequency is close to unity.

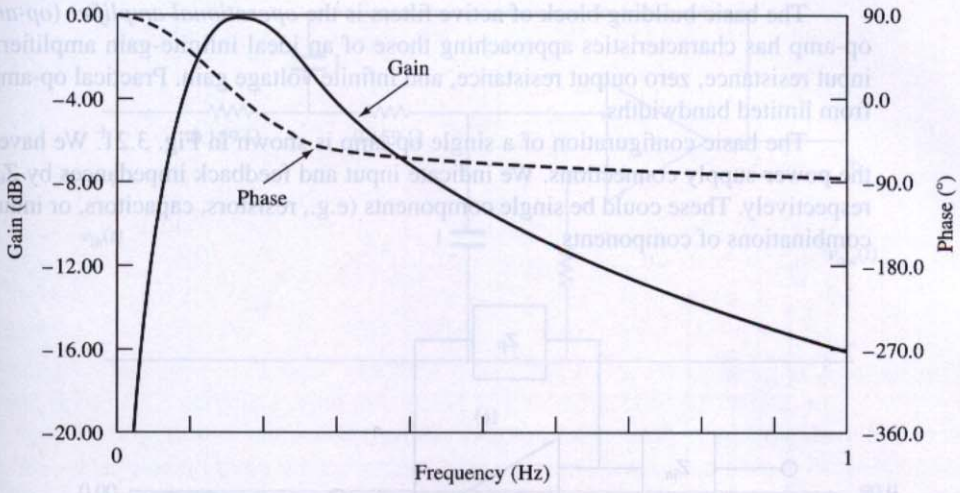


Figure 3.19 Characteristics of *RLC* filter.

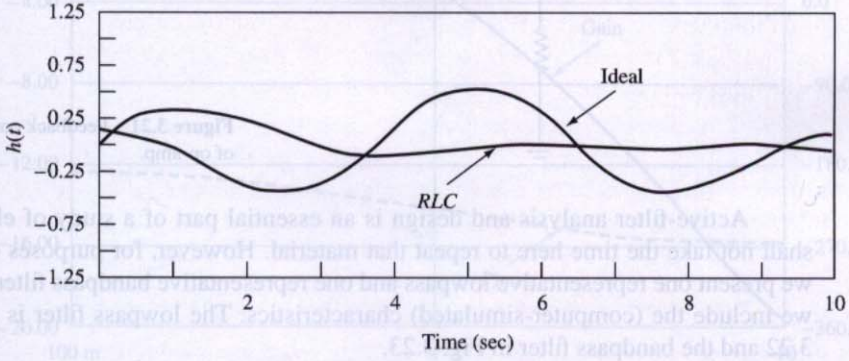


Figure 3.20 Comparison of impulse responses.

As in the case of the lowpass filter, improvements are possible by including additional components. We shall reserve our presentation of the bandpass Butterworth filter to the next section, where we show the active version of this filter.

3.6 ACTIVE FILTERS

In Section 3.5, we examined some simple realizations of filters using inductors, capacitors, and resistors. Such filters are called *passive*, since all component parts either absorb or store energy.

A filter is called *active* if it contains devices that deliver energy to the rest of the circuit. Active filters do not absorb part of the desired signal energy, as do passive filters. They are versatile and simple to design, and arbitrary causal transfer functions can be realized. For some applications, such as audio filtering, the passive filter requires an impractically large number of inductors and capacitors.

The basic building block of active filters is the *operational amplifier* (*op-amp*). The op-amp has characteristics approaching those of an ideal infinite-gain amplifier: infinite input resistance, zero output resistance, and infinite voltage gain. Practical op-amps suffer from limited bandwidths.

The basic configuration of a single op-amp is shown in Fig. 3.21. We have omitted the power-supply connections. We indicate input and feedback impedances by Z_{in} and Z_F, respectively. These could be single components (e.g., resistors, capacitors, or inductors) or combinations of components.

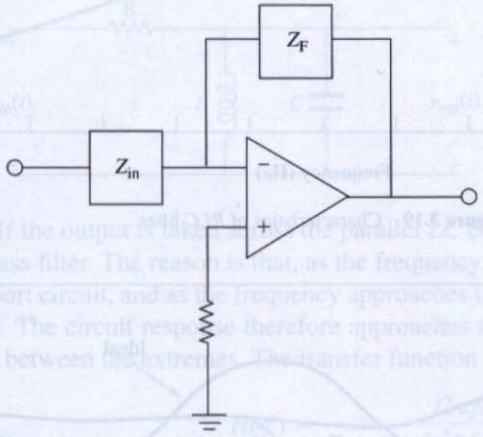


Figure 3.21 Feedback configuration of op-amp.

Active-filter analysis and design is an essential part of a study of electronics. We shall not take the time here to repeat that material. However, for purposes of illustration, we present one representative lowpass and one representative bandpass filter. In each case, we include the (computer-simulated) characteristics. The lowpass filter is shown in Fig. 3.22 and the bandpass filter in Fig. 3.23.

3.7 TIME-BANDWIDTH PRODUCT

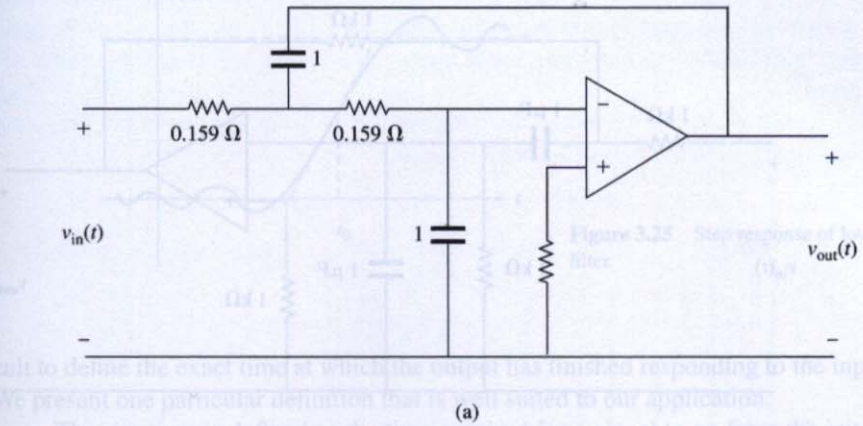
In designing a communication system, an important consideration is the *bandwidth* of the system. The bandwidth is the range of frequencies the system is capable of handling.

The bandwidth is related to the Fourier transform of a function of time. It is not directly definable in terms of the function, unless we use intuitive statements about how quickly the function changes value.

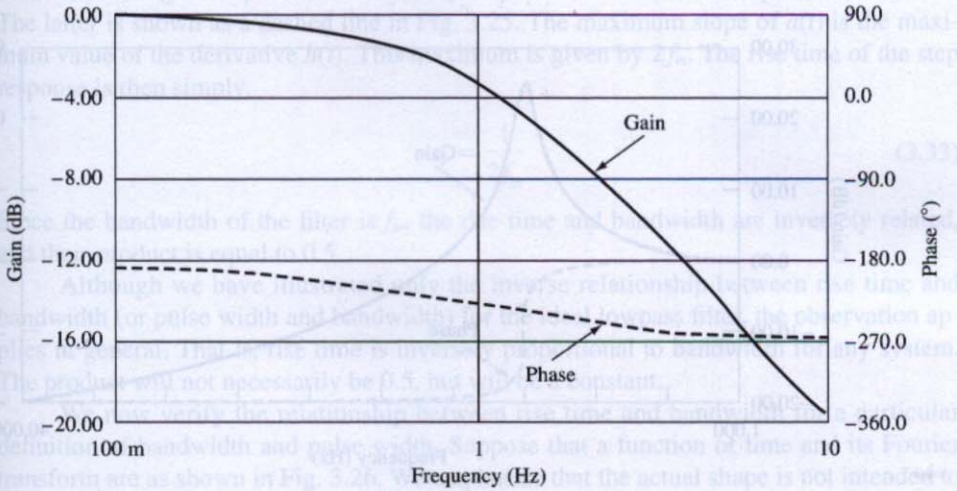
Physical quantities of importance in communication system design include the minimum width of a time pulse and the minimum time in which the output of a system can jump from one level to another. We will show that both of these physical quantities are related to the bandwidth. We start with a specific example and then generalize the result.

The impulse response of the ideal lowpass filter is

$$h(t) = \frac{\sin 2\pi f_m(t - t_0)}{\pi(t - t_0)} \tag{3.32}$$



(a)

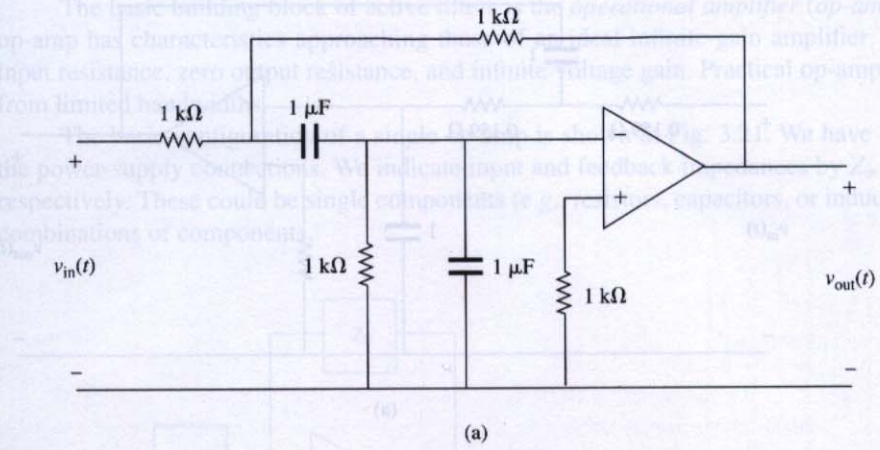


(b)

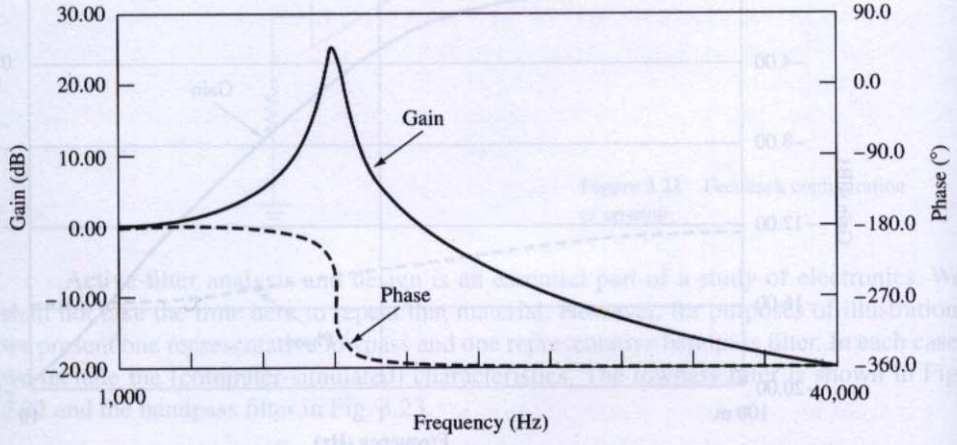
Figure 3.22 Active lowpass filter.

This $h(t)$ and the corresponding $H(f)$ are shown in Fig. 3.24. We use this transform pair to make two observations. First, the width of the largest lobe of $h(t)$ is $1/f_m$. This width is inversely proportional to the bandwidth of the signal. In fact, since the resulting bandwidth (difference between lowest and highest frequency) is f_m, the product of the pulse width and the bandwidth is unity.

The second observation regarding the lowpass filter requires that we find the step response of the filter. Since a step is the time integral of an impulse, and the lowpass filter is a linear system, the step response is the time integral of the impulse response. The step response, $a(t)$, is shown in Fig. 3.25. We now show that the *rise time* of this response is inversely proportional to the bandwidth of the filter. First we must define rise time. There are several common definitions, each of which attempts to mathematically define the length of time it takes the output to respond to a change, or jump, in the input. In practice, it is diffi-



(a)



(b)

Figure 3.23 Active bandpass filter.

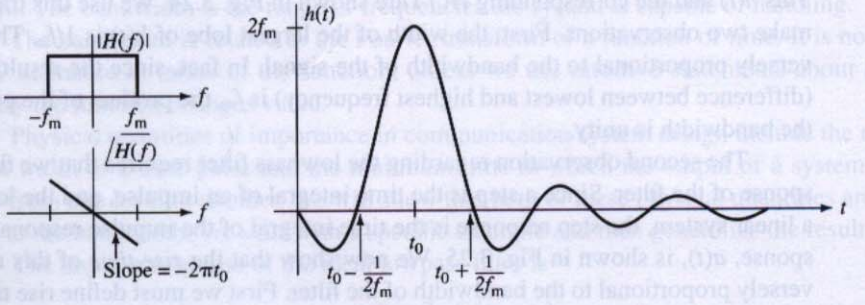


Figure 3.24 Characteristics of ideal lowpass filter.

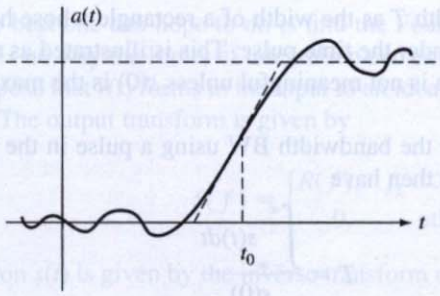


Figure 3.25 Step response of lowpass filter.

cult to define the exact time at which the output has finished responding to the input jump. We present one particular definition that is well suited to our application.

The rise time is defined as the time required for a signal to go from the initial to the final value along a ramp with constant slope equal to the maximum slope of the function. The latter is shown as a dashed line in Fig. 3.25. The maximum slope of $a(t)$ is the maximum value of the derivative $h(t)$. This maximum is given by $2f_m$. The rise time of the step response is then simply

$$t_r = \frac{1}{2f_m} \tag{3.33}$$

Since the bandwidth of the filter is f_m, the rise time and bandwidth are inversely related, and their product is equal to 0.5.

Although we have illustrated only the inverse relationship between rise time and bandwidth (or pulse width and bandwidth) for the ideal lowpass filter, the observation applies in general. That is, rise time is inversely proportional to bandwidth for any system. The product will not necessarily be 0.5, but will be a constant.

We now verify the relationship between rise time and bandwidth for a particular definition of bandwidth and pulse width. Suppose that a function of time and its Fourier transform are as shown in Fig. 3.26. We emphasize that the actual shape is not intended to be that shown in the figure. Indeed, if the pulse is time limited, the transform cannot go identically to zero over any range of frequencies. It is also unrealistic to think that the functions are monotonic. We present the pictures only to help understand the definitions of pulse width and bandwidth.

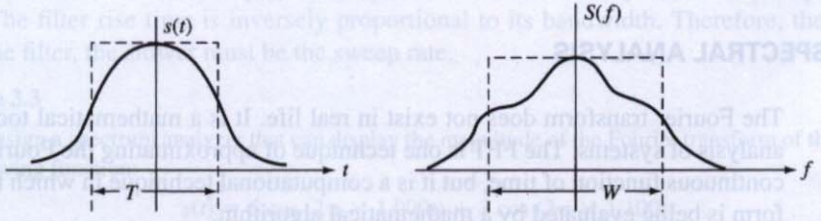


Figure 3.26 Definition of pulse width and bandwidth.

We define the pulse width T as the width of a rectangle whose height matches $s(0)$, and area is the same as that under the time pulse. This is illustrated as a dashed line in the figure. Note that the definition is not meaningful unless $s(0)$ is the maximum of the waveform.

Equivalently, we define the bandwidth BW using a pulse in the frequency domain, as illustrated in the figure. We then have

$$T = \frac{\int_{-\infty}^{\infty} s(t)dt}{s(0)} \qquad \text{BW} = \frac{\int_{-\infty}^{\infty} S(f)df}{S(0)} \tag{3.34}$$

The product of these two is

$$T\cdot\text{BW} = \frac{\int_{-\infty}^{\infty} s(t)dt \int_{-\infty}^{\infty} S(f)df}{s(0)S(0)} \tag{3.35}$$

We now use the Fourier transform integral to find

$$S(0) = \int_{-\infty}^{\infty} s(t)e^{-j2\pi ft}dt\Big|_{f=0} = \int_{-\infty}^{\infty} s(t)dt \tag{3.36}$$

The inverse transform integral is used to find

$$s(0) = \int_{-\infty}^{\infty} S(f)e^{-j2\pi ft}df\Big|_{t=0} = \int_{-\infty}^{\infty} S(f)df \tag{3.37}$$

Substituting Eqs. (3.36) and (3.37) into Eq. (3.35), we find that

$$T\cdot\text{BW} = 1 \tag{3.38}$$

The product of the pulse width with the bandwidth is unity; hence, the two parameters are inversely related.

Clearly, the faster we desire a signal to change from one level to another, the more space on the frequency axis we must allow. This proves significant in digital communication, where the bit transmission rate is limited by the bandwidth of the channel.

3.8 SPECTRAL ANALYSIS

The Fourier transform does not exist in real life. It is a mathematical tool that aids in the analysis of systems. The FFT is one technique of approximating the Fourier transform of a continuous function of time, but it is a computational technique in which the Fourier transform is being evaluated by a mathematical algorithm.

There are severe limitations when one attempts to find the continuous Fourier transform of a time signal using a real analog system. First, since any real system must be

causal, the best one can hope to do is find the Fourier transform based on past input values; there is no way the limits of integration can extend over all time—past and future.

Suppose that $r(t)$ forms in the input to an ideal bandpass filter with $H(f)$ as shown in Fig. 3.27. The output transform is given by

$$S(f) = \begin{cases} R(f), & f_L < |f| < f_H \\ 0, & \text{otherwise} \end{cases} \tag{3.39}$$

The function $s(t)$ is given by the inverse transform of $S(f)$:

$$s(t) = \int_{f_L}^{f_H} R(f)e^{j2\pi ft}df + \int_{-f_H}^{-f_L} R(f)e^{j2\pi ft}df \tag{3.40}$$

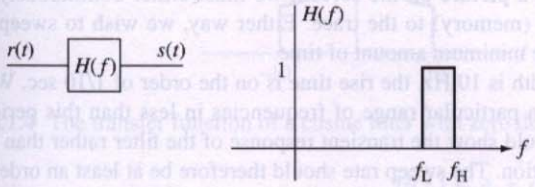


Figure 3.27 Bandpass filter.

If f_H is very close to f_L (i.e., if the filter is *narrowband*), we can assume that the integrand is approximately constant over the entire range of integration. Therefore, if f_{av} is the center of the filter passband, we have

$$s(t) \approx (f_H - f_L)[R(f_{av})e^{j2\pi f_{av}t} + R(-f_{av})e^{-j2\pi f_{av}t} \tag{3.41}$$

Now, since $R(-f_{av}) = R^*(f_{av})$, we have

$$s(t) = (f_H - f_L)|R(f_{av})| \cos[2\pi f_{av}t + \underline{/R(f_{av})}] \tag{3.42}$$

Thus, the magnitude of the output is proportional to the magnitude of the input transform evaluated at f_{av}, and the phase is shifted by the phase of $R(f_{av})$.

In many practical spectrum analyzers, the bandpass filter is swept across a range of frequencies, and the magnitude of the output varies approximately with $|R(f|$.

There are three primary sources of error. First, while the bandpass filter is narrow, its bandwidth is not zero. This affects the *resolution* of the output. Second, the filter is causal and nonideal, which introduces error. Finally, when the center frequency of the filter varies with time (i.e., when it is *swept*), the output does not necessarily reach its steady-state value. The filter rise time is inversely proportional to its bandwidth. Therefore, the narrower the filter, the slower must be the sweep rate.

Example 3.3

Design a spectrum analyzer that can display the magnitude of the Fourier transform of the following function:

$$s(t) = 5\cos(2\pi \times 1{,}000t) + 3\cos(2\pi \times 1{,}100t)$$

Solution: Assuming that we can build a narrowband bandpass filter and sweep it across a range of frequencies (we will see a much better way to do this in Chapter 6), the design of the

spectrum analyzer consists of choosing the frequency range and bandwidth of the filter and also choosing the rate at which the filter sweeps across the range of frequency.

Although we do not know $s(t)$ in advance (if we did, why bother with a spectrum analyzer?), we must assume that we know something about the range of frequencies $s(t)$ occupies and also about the required resolution. In fact, if we wish the approximate transform to look anything like the theoretical transform (two impulses), the bandwidth of the filter would have to be much smaller than the 100-Hz spacing between frequency components. If it were not small enough, the output would consist of components of each of the two frequencies, and we would not be capable of resolving these frequency components. Suppose we choose a filter bandwidth of 10 Hz and a center frequency sweep range of 900 Hz to 1,200 Hz.

We would probably display the result on a monitor. We do this by controlling the vertical displacement with the filter output magnitude and using a ramp generator (time scale) for the horizontal axis. The ramp must be synchronized with the filter sweep rate. The x-axis would then display frequency, while the y-axis would display the magnitude of the transform.

In order to "paint" a picture on the screen, we must either continuously repeat the sweep or add persistence (memory) to the trace. Either way, we wish to sweep across the range of frequencies in the minimum amount of time.

If the filter bandwidth is 10 Hz, the rise time is on the order of 1/10 sec. We therefore would not want to leave a particular range of frequencies in less than this period of time. Otherwise, the display would show the transient response of the filter rather than the Fourier transform magnitude function. The sweep rate should therefore be at least an order of magnitude less than 100 Hz/sec, or 10 Hz/sec. At this rate, it would take 30 seconds to sweep the entire range, so for the desired level of resolution, we would have to use a high-persistence CRT. Figure 3.28 illustrates the resulting spectrum analyzer output.

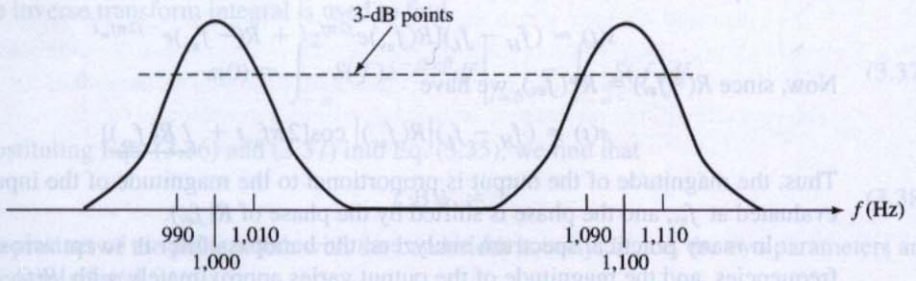


Figure 3.28 Spectrum analyzer output for Example 3.3.

PROBLEMS

3.1.1 You are given a system with input $r(t)$ and output $s(t)$. You are told that when $r(t) = 0$, $s(t)$ is not equal to zero. Show that this system cannot obey superposition.

3.1.2 A filter has the sinusoidal amplitude response shown in Fig. P3.1.2. The phase response is linear with slope $-2\pi t_0$.

(a) Find the system response due to an input $\cos 2\pi t$.

(b) Find the system response due to an input $(\sin 2\pi t)/t$.

(c) Find the system response due to an input $(\sin 20\pi t)/t$.

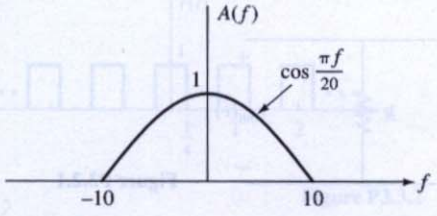


Figure P3.1.2

3.1.3 Repeat Problem 3.1.2 for the amplitude response shown in Fig. P3.1.3.

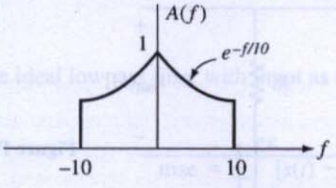


Figure P3.1.3

3.1.4 The transfer function of a cosine filter with zero delay is

$$H(f) = A + a\cos\frac{n\pi f}{f_m}$$

(a) Find the impulse response $h(t)$.
(b) Find the filter output due to an input

$$r(t) = \frac{\sin \pi t}{t}\cos 1{,}000\ \pi t$$

3.1.5 A system is shown in Fig. P3.1.5. Assume that $\theta(f) = -2\pi f t_0$. Expand $A(f)$ in a series in order to find the response due to a unit-amplitude, unit-width square pulse.

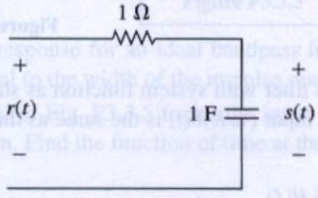


Figure P3.1.5

3.2.1 For the circuit shown in Fig. P3.2.1:
(a) Find $H(f)$.
(b) Plot $|H(f)|$ as a function of frequency.
(c) What function is this circuit performing?

3.2.2 For the circuit shown in Fig. P3.2.2:
(a) Find $H(f)$.
(b) Plot $|H(f)|$ as a function of frequency.
(c) What function is this circuit performing?

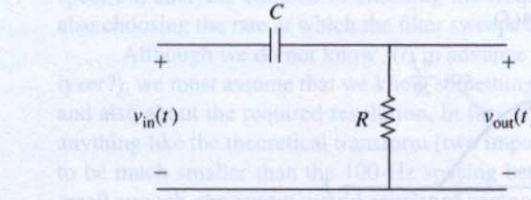


Figure P3.2.1

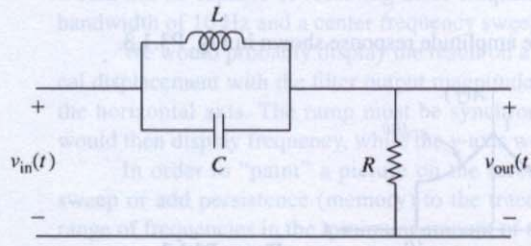


Figure P3.2.2

3.2.3 You are given the system shown in Fig. P3.2.3. The output of the system is $i(t)$. Find the phase distortion when the input is given by

$$r(t) = \frac{\sin t}{t} \cos 200\,t$$

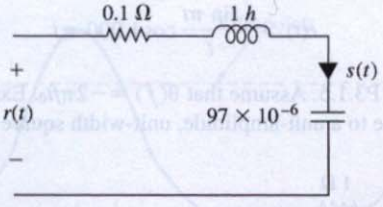


Figure P3.2.3

3.3.1 Consider the ideal lowpass filter with system function as shown in Fig. P3.3.1. Show that the response of this filter to an input $(\pi/K)\delta(t)$ is the same as that to $\sin(Kt)/Kt$.

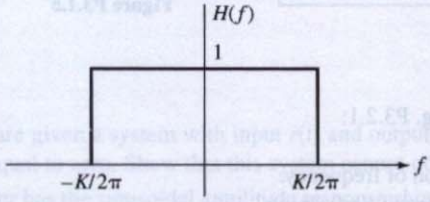


Figure P3.3.1

3.3.2 You are given the ideal lowpass filter with input as shown in Fig. P3.3.2. An error function is defined as the difference between the input and the output; that is,

$$e(t) = r(t) - s(t)$$

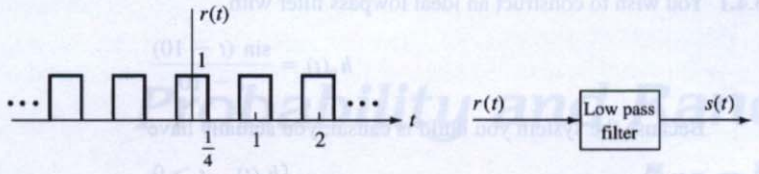


Figure P3.3.2

Find the error function if the filter cutoff frequency is

(a) $f_m = 2.5$ Hz

(b) $f_m = 3.5$ Hz

(c) $f_m = 4.5$ Hz

3.3.3 You are given the ideal lowpass filter with input as shown in Fig. P3.3.3. The mean square error is defined by

$$\text{mse} = \frac{1}{T}\int_0^T [s(t) - r(t)]^2\, dt$$

(a) Show that $s(t)$ is the average value of $r(t)$.

(b) Find the mean square error.

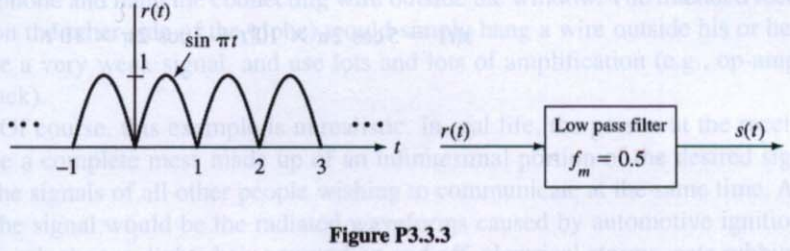


Figure P3.3.3

3.3.4 Find the impulse response for an ideal bandpass filter. Show that the filter bandwidth is inversely proportional to the width of the impulse response.

3.3.5 The periodic signal of Fig. P3.3.5 forms the input to an ideal bandpass filter with amplitude and phase as shown. Find the function of time at the output of the filter.

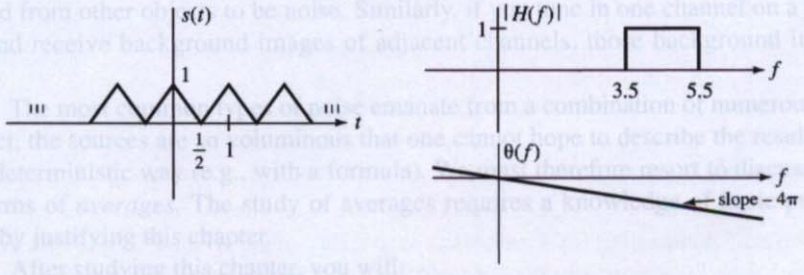


Figure P3.3.5

3.4.1 You wish to construct an ideal lowpass filter with

$$h_0(t) = \frac{\sin\,(t-10)}{t-10}$$

Because the system you build is causal, you actually have

$$h(t) = \begin{cases} h_0(t), & t > 0 \\ 0, & t < 0 \end{cases}$$

(a) Find $H(f)$, and compare it to the system function of the ideal filter.
(b) Find the output when the input is

$$r(t) = \frac{\sin t}{t}$$

(c) Find the error (difference between output and input) for the input of part (b).

3.5.1 Compare the step response of an RC circuit (output taken across the capacitor) with that of an ideal lowpass filter. Find the value of f_m for the ideal filter (in terms of R and C), which minimizes the integrated square error between the two step responses.

3.5.2 Design a Butterworth lowpass filter with a 3-dB cutoff at 500 Hz. The roll-off of the filter must be such that the amplitude response is attenuated by at least 50 dB at a frequency of 3 kHz.

3.8.1 Design a spectrum analyzer that can display the magnitude of the Fourier transform of the function

$$s(t) = 5\cos 2\pi \times 10^6 t + 3\cos 2\pi \times 10^7 t$$

4

Probability and Random Analysis

4.0 PREVIEW

What We Will Cover and Why You Should Care

Studying communication systems without taking noise into consideration is analogous to learning how to drive a car by practicing in a huge abandoned parking lot. While it is a valuable first step, it is not very realistic. Were it not for noise, the communication of signals would be trivial indeed. You could simply build a circuit composed of a battery and a microphone and hang the connecting wire outside the window. The intended receiver (perhaps on the other side of the globe) would simply hang a wire outside his or her window, receive a very weak signal, and use lots and lots of amplification (e.g., op-amps without feedback).

Of course, this example is unrealistic. In real life, the person at the receiver would receive a complete mess made up of an infinitesimal portion of the desired signal mixed with the signals of all other people wishing to communicate at the same time. Also mixed with the signal would be the radiated waveforms caused by automotive ignitions, people dialing telephones, lights being turned on and off, electrical storms, cats rubbing their fur on rugs, sunspots, and a virtual infinity of other spurious signals. The real challenge of communication is separating the desired signal from the undesired junk.

Anything other than the desired signal is called *noise*. Noise includes the effects just described, but it also includes some things you normally would not think of as noise. For example, a radar system attempting to track a particular object considers the signals returned from other objects to be noise. Similarly, if you tune in one channel on a television set and receive background images of adjacent channels, those background images are noise.

The most common types of noise emanate from a combination of numerous sources. In fact, the sources are so voluminous that one cannot hope to describe the resulting noise in a deterministic way (e.g., with a formula). We must therefore resort to discussing noise in terms of *averages*. The study of averages requires a knowledge of basic probability, thereby justifying this chapter.

After studying this chapter, you will:

- understand the basics of probability theory
- be able to solve problems involving random quantities

- have the tools necessary to evaluate the performance of communication systems in the presence of noise
- understand the matched filter, which is a building block in digital receivers.

Necessary Background

To understand basic probability, you need to know only elementary calculus. To understand random processes (which are discussed later in the chapter), you also need to know basic system theory.

4.1 BASIC ELEMENTS OF PROBABILITY THEORY

Probability theory can be approached either using theoretical mathematics or through empirical reasoning. The *mathematical approach* embeds probability theory within a study of *abstract set theory.* In contrast, the *empirical approach* satisfies one's intuitions. In our basic study of communication, we will find the empirical approach to be sufficient, although advanced study and references to current literature require extending these concepts using principles of set theory.

Before we define probability, we must extend our vocabulary by defining some other important terms:

An *experiment* is a set of rules governing an operation that is performed.
An *outcome* is the result realized after performing an experiment one time.
An *event* is a combination of outcomes.

Consider the experiment defined by flipping a single die (half of a pair of dice) and observing which of the six faces is at the top when the die comes to rest. (Notice how precise we are being: If you simply say "flipping a die," you could mean that you observe the *time* at which it hits the floor.) There are six possible outcomes, namely, any one of the six surfaces of the die facing upward after the performance of the experiment.

There are many possible events (64, to be precise). One event would be that of "an even number of dots showing." This event is a combination of the three outcomes of two dots, four dots, and six dots showing. Another event is "one dot showing." This event is known as an *elementary event,* since it is the same as one of the outcomes. Of the 64 possible events, six represent elementary events. You should be able to list the 64 events. Try it! If you come up with only 62 or 63, you are probably missing the combination of all outcomes and/or the combination of no outcomes.

4.1.1 Probability

We now define what is meant by the probability of an event. Suppose that an experiment is performed N times, where N is very large. Furthermore, suppose that in n of these N experiments, the outcome belongs to an event A (e.g., consider flipping a die 1,000 times, and in 495 of these flips the outcome is "even"; then $N = 1{,}000$ and $n = 495$). If N is large

enough, the probability of event A is given by the ratio n/N. That is, the probability is the fraction of times that the event occurs. Formally, we define the probability of event A as

$$\Pr\{A\} = \lim_{N\to\infty} \frac{n_A}{N} \tag{4.1}$$

In Eq. (4.1), n_A is the number of times that the event A occurs in N performances of the experiment. This definition is intuitively satisfying. For example, if a coin were flipped many times, the ratio of the number of heads to the total number of flips would approach $\frac{1}{2}$. We therefore define the probability of a head to be $\frac{1}{2}$. This simple example shows why N must approach infinity. Suppose, for example, you flipped a coin three times and in two of these flips, the outcome were heads. You would certainly not be correct in assuming the probability of heads to be $\frac{2}{3}$!

Suppose that we now consider two different events, A and B, with probabilities

$$\Pr\{A\} = \lim_{N\to\infty} \frac{n_A}{N} \quad \text{and} \quad \Pr\{B\} = \lim_{N\to\infty} \frac{n_B}{N} \tag{4.2}$$

If A and B could not possibly occur at the same time, we call them *disjoint.* For example, the events "an even number of dots" and "two dots" are not disjoint in the die-throwing example, while the events "an even number of dots" and "an odd number of dots" are disjoint.

The probability of event A *or* event B is the number of times A or B occurs divided by N. If A and B are disjoint, this is

$$\Pr\{A \text{ or } B\} = \lim_{N\to\infty} \frac{n_A + n_B}{N} = \Pr\{A\} + \Pr\{B\} \tag{4.3}$$

Equation (4.3) expresses the additivity concept: If two events are disjoint, the probability of their sum is the sum of their probabilities.

Since each of the outcomes (elementary events) is disjoint from every other outcome, and each event is a sum of outcomes, it would be sufficient to assign probabilities only to the elementary events. We could derive the probability of any event from these given probabilities. For example, in the die-flipping experiment, the probability of an even outcome is the sum of the probabilities of "2 dots," "4 dots," and "6 dots."

Example 4.1

Consider the experiment of flipping a coin twice and observing which side is facing up when the coin comes to rest. List the outcomes, the events, and their respective probabilities.

Solution: The outcomes of this experiment are (letting H denote heads and T tails)

HH, HT, TH, and *TT*

We shall assume that somebody has used intuitive reasoning or has performed this experiment enough times to establish that the probability of each of the four outcomes is $\frac{1}{4}$. There are 16 events, of combinations of these outcomes:

$\{HH\}, \{HT\}, \{TH\}, \{TT\}$
$\{HH,HT\}, \{HH,TH\}, \{HH,TT\}, \{HT,TH\}, \{HT,TT\}, \{TH,TT\}$
$\{HH,HT,TH\}, \{HH,HT,TT\}, \{HH,TH,TT\}, \{HT,TH,TT\}$
$\{HH,HT,TH,TT\}$, and $\{\phi\}$

Note that the comma within the braces is read "or." Thus, the events {HH,HT} and {HT,HH} are identical, and we list this event only once. For completeness, we have included the zero event, denoted $\{\phi\}$. This is the event made up of none of the outcomes and is called the *null* event. We also include the event consisting of all of the outcomes, the so-called *certain* event.

Using the additivity rule, the probability of each of these events is the sum of the probabilities of the outcomes comprising each event. Therefore,

$$\Pr\{HH\} = \Pr\{HT\} = \Pr\{TH\} = \Pr\{TT\} = \tfrac{1}{4}$$
$$\Pr\{HH,HT\} = \Pr\{HH,TH\} = \Pr\{HH,TT\} = \Pr\{HT,TH\} = \Pr\{HT,TT\} = \Pr\{TH,TT\} = \tfrac{1}{2}$$
$$\Pr\{HH,HT,TH\} = \Pr\{HH,HT,TT\} = \Pr\{HH,TH,TT\} = \Pr\{HT,TH,TT\} = \tfrac{3}{4}$$
$$\Pr\{HH,HT,TH,TT\} = 1$$
$$\Pr\{\phi\} = 0$$

The next-to-last probability indicates that the event made up of all four outcomes is the *certain* event. It has probability 1 of occurring, since each time the experiment is performed, the outcome must belong to this event. Similarly, the null event (the last probability) has probability zero of occurring, since each time the experiment is performed, the outcome does not belong to the zero event.

4.1.2 Conditional Probabilities

We would like to be able to tell whether one random quantity has any effect on another. For instance, in the die experiment, if we knew the time at which the die hit the floor, would it tell us anything about which face was showing? In a more practical case, if we knew the frequency of a random noise signal, would this tell us anything about its amplitude? These questions lead naturally into a discussion of *conditional probabilities.*

Let us examine two events, A and B. The probability of event A *given that event* B *has occurred* is defined by

$$\Pr\{A/B\} = \frac{\Pr\{A \text{ AND } B\}}{\Pr\{B\}} \tag{4.4}$$

For example, if A represented two dots appearing in the die experiment and B represented an even number of dots, the probability of A given B would be the probability of two dots appearing, assuming that we know the outcome is either two, four, or six dots appearing. Thus, the conditional statement has reduced the scope of possible outcomes from six to three. We would intuitively expect the answer to be $\frac{1}{3}$. Now, from Eq. (4.4), the probability of "A AND B" is the probability of getting two *AND* an even number of dots simultaneously. (In set theory, this is known as the *intersection.*) It is simply the probability of two dots appearing, or $\frac{1}{6}$. The probability of B is the probability of two, four, or six dots appearing, which is $\frac{1}{2}$. The ratio is $\frac{1}{3}$, as expected.

Similarly, we could have defined event A as "an even number of dots" and event B as "an odd number of dots." The event "A AND B" would then be the zero event, and $\Pr\{A/B\}$ would be zero. This is reasonable because the probability of an even outcome assuming that an odd outcome occurred is clearly zero.

Two events, A and B, are said to be *independent* if

$$\Pr\{A/B\} = \Pr\{A\} \tag{4.5}$$

Thus, if A and B are independent, the probability of A given that B occurred is simply the probability of A. Knowing that B has occurred tells nothing about A. Plugging Eq. (4.5) into Eq. (4.4) shows that if A and B are independent, then

$$\Pr\{A \text{ and } B\} = \Pr\{A\}\Pr\{B\} \tag{4.6}$$

You have probably used this fact before in simple experiments. For example, we assumed that the probability of flipping a coin and having it land with heads facing up was $\frac{1}{2}$. Hence, the probability of flipping the coin twice and getting two heads is $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$. This is true because the events are independent of each other.

Example 4.2

A coin is flipped twice. The following four different events are defined:

A is the event of getting a head on the first flip.
B is the event of getting a tail on the second flip.
C is the event of getting a match between the two flips.
D is the elementary event of getting a head on both flips.

(a) Find $\Pr\{A\}$, $\Pr\{B\}$, $\Pr\{C\}$, $\Pr\{D\}$, $\Pr\{A/B\}$, and $\Pr\{C/D\}$.
(b) Are A and B independent? Are C and D independent?

Solution: **(a)** The events are defined by the following combination of outcomes:

$$A = \{HH, HT\}$$
$$B = \{HT, TT\}$$
$$C = \{HH, TT\}$$
$$D = \{HH\}$$

Therefore,

$$\Pr\{A\} = \Pr\{B\} = \Pr\{C\} = \tfrac{1}{2}$$
$$\Pr\{D\} = \tfrac{1}{4}$$

(b) To find $\Pr\{A/B\}$ and $\Pr\{C/D\}$, we use Eq. (4.4):

$$\Pr\{A/B\} = \frac{\Pr\{A \text{ AND } B\}}{\Pr\{B\}}$$
$$\Pr\{C/D\} = \frac{\Pr\{C \text{ AND } C\}}{\Pr\{D\}}$$

The event $\{A \text{ AND } B\}$ is $\{HT\}$. The event $\{C \text{ AND } D\}$ is $\{HH\}$. Therefore,

$$\Pr\{A/B\} = \frac{\frac{1}{4}}{\frac{1}{2}} = 0.5$$

$$\Pr\{C/D\} = \frac{\frac{1}{4}}{\frac{1}{4}} = 1$$

Since $\Pr\{A/B\} = \Pr\{A\}$, the event of a head on the first flip is independent of that of a tail on the second flip. Since $\Pr\{C/D\} \neq \Pr\{C\}$, the event of a match and that of two heads are not independent.

4.1.3 Random Variables

We would like to perform several forms of analysis on probabilities. It is not too satisfying to work with symbols such as "heads," "tails," and "two dots." It would be preferable to work with numbers. We therefore associate a real number with each possible outcome of an experiment. For example, in the single-flip-of-the-coin experiment, we could associate the number 0 with "tails" and 1 with "heads." We could just as well (although we won't) associate π with "heads" and 207 with "tails."

The mapping (function) that assigns a number to each outcome is called a *random variable.*

Once a random variable is assigned, we can perform many forms of analysis. We can, for example, plot the various outcome probabilities as a function of the random variable. An extension of that type of plot is the *distribution function* $F(x)$. If the random variable is denoted by[1] X, then the distribution function $F(x_0)$ is defined by

$$F(x_0) = \Pr\{X \leq x_0\} \tag{4.7}$$

We note that $\{X \leq x_0\}$ defines an event, or combination of outcomes.

Example 4.3

Assign two different random variables to the "one-flip-of-the-die" experiment, and plot the two resulting distribution functions.

Solution: The first assignment we will choose is the one that is naturally suggested by this particular experiment. That is, we assign the number 1 to the outcome described by the face with one dot facing up, we assign the number 2 to "two dots," 3 to "three dots," and so on. We therefore see that the event $\{X \leq X_0\}$ includes the one-dot outcome if x_0 is between 1 and 2. If x_0 is between 2 and 3, the event includes the one-dot and two-dot outcomes. Thus, the distribution function is as shown in Fig. 4.1(a).

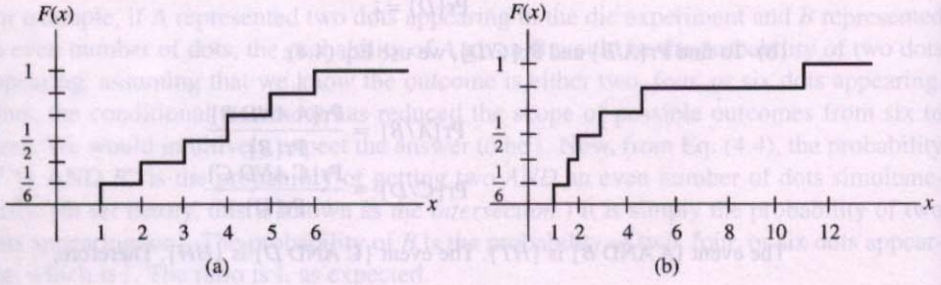


Figure 4.1 Distribution function for Example 4.3.

[1]We shall use uppercase letters for random variables and lowercase letters for the values they can take on. Thus, $X = x_0$ means that the random variable X is equal to the number x_0.

Let us now choose a different assignment of the random variable, one representing a less natural choice:

Outcome	Random Variable
One dot	1
Two dots	π
Three dots	2
Four dots	$\sqrt{2}$
Five dots	11
Six dots	5

We have chosen strange numbers to illustrate that the mapping is arbitrary. The resulting distribution function is plotted as Fig. 4.1(b). As an example, let us verify one point on the distribution function, the point for $x = 3$. The event $\{X \leq 3\}$ is the event made up of the following three outcomes: one dot, three dots, and four dots. This is true because the value of the random variable assigned to each of these three outcomes is less than 3.

A distribution function can never decrease with increasing argument. The reason is that an increase in argument can only add outcomes to the event, and the probabilities of these added outcomes cannot be negative. We also easily verify that

$$F(-\infty) = 0 \quad \text{and} \quad F(+\infty) = 1 \tag{4.8}$$

4.1.4 Probability Density Function

The *probability density function* is defined as the derivative of the distribution function. Using the symbol $p_X(x)$ for the density, we have

$$p_X(x) = \frac{dF(x)}{dx} \tag{4.9}$$

Since $p(x)$ is the derivative of $F(x)$, $F(x)$ is the integral of $p(x)$:

$$F(x_0) = \int_{-\infty}^{x_0} p_X(x)\, dx \tag{4.10}$$

The random variable can be used to define any event. For example, $\{x_1 < X \leq x_2\}$ defines an event. Since the events $\{X \leq x_1\}$ and $\{x_1 < X \leq x_2\}$ are disjoint, the additivity principle can be used to prove that

$$\Pr\{X \leq x_1\} + \Pr\{x_1 < X \leq x_2\} = \Pr\{X \leq x_2\}$$

or

$$\Pr\{x_1 < X \leq x_2\} = \Pr\{X \leq x_2\} - \Pr\{X \leq x_1\} \tag{4.11}$$

Combining Eqs. (4.10) and (4.11), we have the important result,

$$\Pr\{x_1 < X \leq x_2\} = \int_{-\infty}^{x_2} P_X(x)dx - \int_{-\infty}^{x_1} P_X(x)dx \tag{4.12}$$

$$= \int_{x_1}^{x_2} P_X(x)dx$$

We now see why $p_X(x)$ is called a density function: The probability that X is between any two limits is given by the area under the density function between these two limits.

Since the distribution function can never decrease with increasing argument, its slope, the density function, can never be negative.[2] Also, since the distribution function approaches unity as its argument approaches infinity, the integral of the density function over infinite limits must be unity.

The examples given previously (die and coin) result in density functions that contain impulses. The random variables associated with such experiments are known as *discrete random variables.* Another class of experiments gives rise to random variables with continuous density functions. This is logically called the class of *continuous random variables.* We present several frequently occurring continuous random variable density functions in Section 4.2. For now, we examine the simplest of these functions, the *uniform density function.* This function is shown in Fig. 4.2(a), where a and b are specified parameters. The height of the density must be such that the total area is unity.

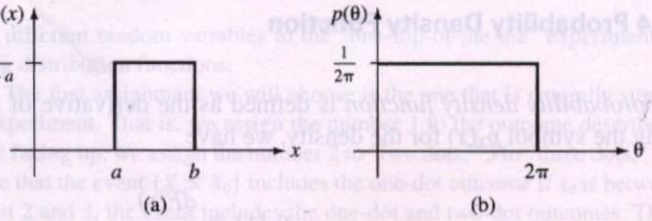


Figure 4.2 Uniform density function.

Let us look at one practical experiment that results in a uniformly distributed random variable. Suppose you were asked to turn on a sinusoidal generator. The output of the generator would be of the form

$$v(t) = A\cos(2\pi f_0 t + \theta) \tag{4.13}$$

Since the absolute time at which you turn on the generator is random, it would be reasonable to expect that θ is uniformly distributed between 0 and 2π. (This is true provided that f_0 is much larger than the reciprocal of your reaction time.) It would thus follow the density function shown in Fig. 4.2(b).

[2]We could infer the same conclusion from Eq. (4.12). If the probability density function were negative over any range of values, we could integrate the curve over that range to get a negative result. This would imply that the probability of the variable being in that range is negative, but that is impossible.

Example 4.4

A random variable is uniformly distributed between 1 and 3. Find the probability that the variable is in the range between 1.5 and 2.

Solution: The density function is as shown in Fig. 4.2(a), where a is 1 and b is 3. In order for this to integrate to unity, the height of the density must be $\frac{1}{2}$. The probability that the variable is between 1.5 and 2 is simply the integral under the curve between these limits. Clearly, this is equal to $\frac{1}{4}$.

4.1.5 Expected Values

Expected values, or averages, are important in communication. The average of the square of a voltage is closely related to the power associated with that voltage. The power of a noise voltage is an important measure of the level of disturbance caused by the voltage.

The expected values that come up often enough to be given names are the *mean*, *variance*, and *moments* of a random variable. We define these terms in this subsection.

Picture yourself as a professor who has just given an examination. How would you average the resulting grades? You would probably add them all together and divide by the number of grades. If an experiment is performed many times, the average of the random variable that results would be found in the same way.

An alternative way to find the sum of grades is to take 100 multiplied by the number of students who got 100 as a grade, add this to 99 times the number of students who got 99, and continue this process for all possible grades. Then divide the sum by the total number of grades. Let us formalize this approach.

Let x_i, $i = 1, 2, \ldots, M$, represent the possible values of the random variable, and let n_i represent the number of times the outcome associated with x_i occurs. Then the average of the random variable after N performances of the experiment is

$$X_{\text{avg}} = \frac{1}{N}\sum_i n_i x_i = \sum_i \frac{n_i}{N} x_i \tag{4.14}$$

Since x_i ranges over all possible values of the random variable,

$$\sum_i n_i = N \tag{4.15}$$

As N approaches infinity, n_i/N becomes the probability, $\Pr\{x_i\}$. Therefore,

$$X_{\text{avg}} = \sum_i x_i \Pr\{x_i\} \tag{4.16}$$

This average value is known as the *mean, expected value,* or *first moment* of X and is given the symbol $E\{x\}$, X_{avg}, m_x, or $\bar{x}$. The words "expected value" should not be taken too literally, since they do not always lend themselves to an intuitive definition. As an example, suppose we assign 1 to heads and 0 to tails in the coin flip experiment. Then the expected value of the random variable is $\frac{1}{2}$. However, no matter how many times you perform the experiment, you will never obtain an outcome with an associated random variable of $\frac{1}{2}$.

Now suppose that we wish to find the average value of a continuous random variable. We can use Eq. (4.16) if we first round off the continuous variable to the nearest mul-

tiple of Δx. Thus, if X is between $(k - \frac{1}{2})\Delta x$ and $(k + \frac{1}{2})\Delta x$, we round it off to $k\Delta x$. The probability of X being in this range is given by the integral of the probability density function:

$$\Pr\left\{\left(k - \frac{1}{2}\right)\Delta x < x \leq \left(k + \frac{1}{2}\right)\Delta x\right\} = \int_{\left(k - \frac{1}{2}\right)\Delta x}^{\left(k + \frac{1}{2}\right)\Delta x} p_X(x)\, dx \tag{4.17}$$

If Δx is small, this can be approximated by $p_X(k\Delta x)\Delta x$. Therefore, Eq. (4.16) can be rewritten as

$$X_{\text{avg}} = \sum_{k=-\infty}^{\infty} k\Delta x p_X(k\Delta x)\Delta x \tag{4.18}$$

As Δx approaches zero, this becomes

$$X_{\text{avg}} = m_x = \int_{-\infty}^{\infty} x p_X(x) dx \tag{4.19}$$

Equation (4.19) is very important. It tells us that to find the average value of x, we simply weight x by the density function and integrate the product.

The same approach can be used to find the average of any function of a random variable. For example, suppose again that you are a professor who gave an exam, but instead of entering the raw percentage score in your grade book, you enter some function of this score, such as e^x, where x is the raw score. If you now wish to average the entries in the book, you would follow the reasoning used earlier [Eqs. (4.14) through (4.19)], with the result that $xp_X(x)$ in Eq. (4.19) gets replaced by $e^x p_X(x)$.

In general, if $y = g(x)$, the expected value of y is given by

$$Y_{\text{avg}} = [g(x)]_{\text{avg}} = \int_{-\infty}^{\infty} g(x) p_X(x) dx \tag{4.20}$$

Equation (4.20) is extremely significant and useful. It tells us that in order to find the expected value of a function of x, we simply integrate that function weighted by the *density of* X. It is not necessary to find the density of the new random variable first.

We often seek the expected value of the random variable raised to a power. This is given the name *moment*. Thus, the expected value of x^n is known as the *nth moment* of the random variable X.

If we first shift the random variable by its mean and then take a moment of the resulting shifted variable, the *central moment* results. Thus, the nth central moment is given by the expected value of $(x - m_x)^n$.

The *second central moment* is extremely important, because it is related to power. It is given the name *variance* and the symbol σ^2. Thus, the variance is

$$\sigma^2 = E\{(x - m_x)^2\} = \int_{-\infty}^{\infty} (x - m_x)^2 p_X(x)\, dx \tag{4.21}$$

The variance is a measure of how far we can expect the variable to deviate from its mean value. As the variance gets larger, the density function tends to "spread out." The square root of the variance, σ, is known as the *standard deviation*.

Example 4.5

Suppose X is uniformly distributed as shown in Fig. 4.3. Find $E\{x\}$, $E\{x^2\}$, $E\{\cos x\}$ and $E\{(x - m_x)^2\}$.

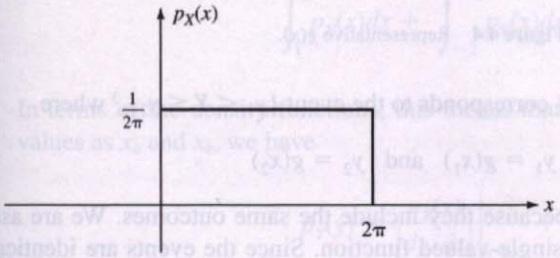


Figure 4.3 Density of x for Example 4.5.

Solution: We apply Eq. (4.20) to find

$$E\{x\} = \int_{-\infty}^{\infty} x p_X(x) dx = \frac{1}{2\pi} \int_0^{2\pi} x\, dx = \pi$$

$$E\{x^2\} = \int_{-\infty}^{\infty} x^2 p_X(x) dx = \frac{1}{2\pi} \int_0^{2\pi} x^2\, dx = \frac{4}{3}\pi^2$$

$$E\{\cos x\} = \int_{-\infty}^{\infty} \cos x\, p_X(x) dx = \frac{1}{2\pi} \int_0^{2\pi} \cos x\, dx = 0$$

$$E\{(x - \pi)^2\} = \int_{-\infty}^{\infty} (x - \pi)^2 p_X(x) dx = \frac{1}{2\pi} \int_0^{2\pi} (x - \pi)^2 dx = \frac{\pi^2}{3}$$

4.1.6 Functions of a Random Variable

"Everybody talks about the weather, but nobody does anything about it." As communication engineers, we ourselves would be open to the same type of criticism if all we ever did was make statements such as "There is a 42-percent probability that the noise will be annoying." A significant part of communication engineering involves changing noise from one form to another in the hope that the new form will be less annoying than the old. We must therefore study the effects of processing on random phenomena.

Consider a function of a random variable, $y = g(x)$, where X is a random variable with known density function. A representative function is shown in Fig. 4.4(a). Since X is random, Y is also random. We are interested in finding the density function of Y.

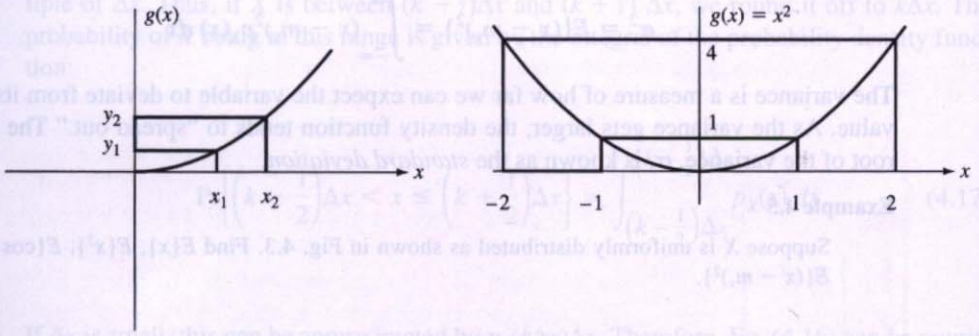


Figure 4.4 Representative $g(x)$.

The event $\{x_1 < X \leq x_2\}$ corresponds to the event $\{y_1 < Y \leq y_2\}$,[3] where

$$y_1 = g(x_1) \quad \text{and} \quad y_2 = g(x_2)$$

The two events are identical because they include the same outcomes. We are assuming for the moment that $g(x)$ is a single-valued function. Since the events are identical, their probabilities must also be equal. That is,

$$\Pr\{x_1 < X \leq x_2\} = \Pr\{y_1 < Y \leq y_2\} \tag{4.22}$$

and in terms of the densities,

$$\int_{x_1}^{x_2} p_X(x)dx = \int_{y_1}^{y_2} p_Y(y)dy \tag{4.23}$$

If we now let x_2 get very close to x_1, then in the limit, Eq. (4.23) becomes

$$p_X(x_1)dx = p_Y(y_1)dy \tag{4.24}$$

and lastly,

$$p_Y(y_1) = \frac{p_X(x_1)}{dy/dx} \tag{4.25}$$

If $y_1 > y_2$, the slope of the curve is negative, and we would find (you should prove this result) that

$$p_Y(y_1) = -\frac{p_X(x_1)}{dy/dx} \tag{4.26}$$

We can account for both of these cases by writing

$$p_Y(y_1) = \frac{p_X(x_1)}{|dy/dx|} \tag{4.27}$$

[3]We are assuming that y1 is less than y2 if x1 is less than x2. That is, g(x) is monotonically increasing and has a positive derivative. If this is not the case, the inequalities would have to be reversed.

Finally, writing $x_1 = g^{-1}(y_1)$, and realizing that y_1 can be a variable (i.e., replace it with y), we have

$$p_Y(y) = \frac{p_X[g^{-1}(y)]}{|dy/dx|} \tag{4.28}$$

If the function $g(x)$ is not monotonic, the event $\{y_1 < Y < y_2\}$ can correspond to more than one interval of the variable X. For example, if $g(x) = x^2$, then the event $\{1 < Y \leq 4\}$ is the same as the event $\{1 < X \leq 2\}$ or $\{-2 < X \leq -1\}$. This is shown in Fig. 4.4(b). Therefore,

$$\int_1^2 p_X(x)dx + \int_{-2}^{-1} p_X(x)dx = \int_1^4 p_Y(y)dy \tag{4.29}$$

In terms of the density functions, this means that $g^{-1}(y)$ has two values. Denoting these values as x_a and x_b, we have

$$p_Y(y) = \frac{p_X(x)}{|dy/dx|}\bigg|_{x=x_a} + \frac{p_X(x)}{|dy/dx|}\bigg|_{x=x_b} \tag{4.30}$$

Example 4.6

A random voltage v is put through a full-wave rectifier. The input voltage is uniformly distributed between -2 volts and $+2$ volts. Find the density of the output of the full-wave rectifier.

Solution: Calling the output y, we have $y = g(v)$, where $g(v)$ and the density of V are sketched in Fig. 4.5. Note that we have let the random variable be equal to the value of voltage.

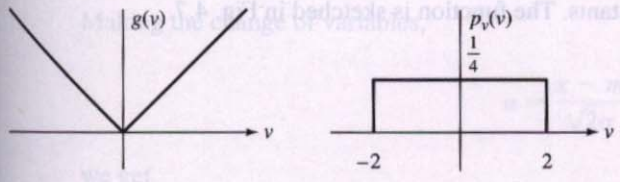


Figure 4.5 $g(v)$ and $p(v)$ for Example 4.6.

At every value of V, $|dg/dv| = 1$. For $y > 0$, $g^{-1}(y) = \pm y$. For $y < 0$, $g^{-1}(y)$ is undefined. That is, there are no values of v for which $g(v)$ is negative. Equation (4.30) is then used to find

$$p_Y(y) = p_V(y) + p_V(-y) \qquad y > 0$$

$$p_Y(y) = 0 \qquad y < 0$$

This result is shown in Fig. 4.6.

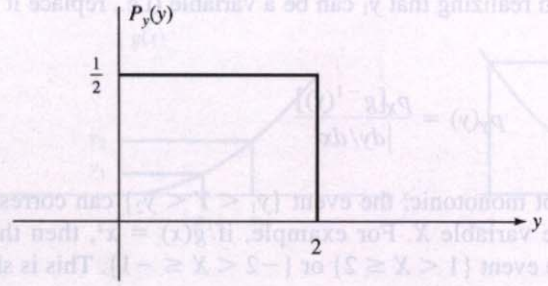


Figure 4.6 Density of output for Example 4.6.

4.2 FREQUENTLY ENCOUNTERED DENSITY FUNCTIONS

We introduced the basic concepts of probability in Section 4.1. The uniform density function was presented. While some experiments in communication lead to this density, the majority of random variables we encounter follow densities other than uniform. The current section explores several of the most frequently encountered densities.

4.2.1 Gaussian Random Variables

The most common density confronted in the real world is called the *Gaussian (or normal) density function*. The reason it is so common is attributed to the *central limit theorem*, a theorem we shall discuss in a few moments. The Gaussian density function is defined by the equation

$$p_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-(x-m)^2}{2\sigma^2}\right] \tag{4.31}$$

where m and σ are given constants. The function is sketched in Fig. 4.7.

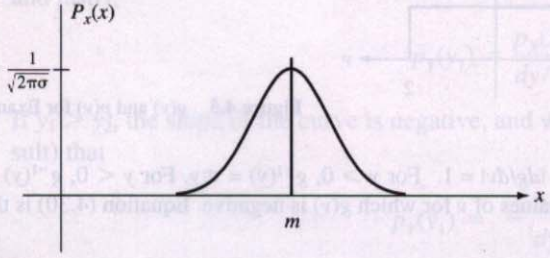


Figure 4.7 Gaussian density function.

The parameter m dictates the center position, or symmetry point, of the density. Evaluating the integral of $xp_X(x)$ would show that m is the mean value of the variable. The other parameter, σ, indicates the spread of the density. Evaluating the integral of $(x-m)^2 p_X(x)$ would show that σ^2 is the variance of the variable, so σ is the standard deviation. As σ increases, the bell-shaped curve gets wider and the peak decreases. Alterna-

tively, as σ decreases, the density sharpens into a narrow pulse with a higher peak. (The area must always be unity.)

To evaluate probabilities that Gaussian variables are within certain ranges, we find it necessary to integrate the density. However, Eq. (4.31) cannot be integrated in closed form, although software such as *Mathcad* can easily be used. The Gaussian density is sufficiently important that this integral has been computed and tabulated under the names *error function* (*erf*) and Q-*function*. The error function is defined as

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-u^2}\, du \tag{4.32}$$

It can be shown that $\operatorname{erf}(\infty) = 1$. Therefore,

$$\begin{aligned} \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-u^2} du &= \frac{2}{\sqrt{\pi}} \int_0^\infty e^{-u^2} du - \frac{2}{\sqrt{\pi}} \int_0^x e^{-u^2} du \\ &= \operatorname{erf}(\infty) - \operatorname{erf}(x) = 1 - \operatorname{erf}(x) \end{aligned} \tag{4.33}$$

For convenience, this last expression is tabulated under the name *complementary error function* (*erfc*). Thus, the relationship between the error function and the complementary error function is

$$\operatorname{erfc}(x) = 1 - \operatorname{erf}(x) \tag{4.34}$$

Both the error function and the complementary error function are tabulated in Appendix III. The area under a Gaussian density with any values of m and σ can be expressed in terms of error functions. For example, the probability that X is between x_1 and x_2 is

$$\Pr\{x_1 < X \le x_2\} = \frac{1}{\sqrt{2\pi}\sigma} \int_{x_1}^{x_2} \exp\left[\frac{-(x-m)^2}{2\sigma^2}\right] dx \tag{4.35}$$

Making the change of variables,

$$u = \frac{x - m}{\sqrt{2}\sigma} \tag{4.36}$$

we get

$$\begin{aligned} \Pr\{x_1 < X \le x_2\} &= \frac{1}{\sqrt{\pi}} \int_{\frac{x_1 - m}{\sqrt{2}\sigma}}^{\frac{x_2 - m}{\sqrt{2}\sigma}} e^{-u^2} du \\ &= \frac{1}{2} \operatorname{erf}\left(\frac{x_2 - m}{\sqrt{2}\sigma}\right) - \frac{1}{2} \operatorname{erf}\left(\frac{x_1 - m}{\sqrt{2}\sigma}\right) \end{aligned} \tag{4.37}$$

We have assumed that both x_1 and x_2 are greater than m, since the error function is not defined for negative arguments. Example 4.7 will deal with a situation where this assumption is not valid.

A companion to the error function is the Q-function. It is sometimes called the complementary error function, or *co-error function.* However, the Q-function differs from the complementary error function of Eq. (4.34) by a constant multiplier and a scaling factor. The Q-function is defined as

$$Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du \tag{4.38}$$

The integrand is a unit-variance, zero-mean Gaussian density function. Note that

$$Q(-\infty) = 1$$
$$Q(-x) = 1 - Q(x)$$

We can relate the Q-function to the error function by making a change of variables:

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-u^2/2}\, du = \frac{1}{\sqrt{\pi}} \int_{x/\sqrt{2}}^\infty e^{-v^2}\, dv$$
$$= \frac{1}{2} \operatorname{erfc}\left(\frac{x}{\sqrt{2}}\right) \tag{4.39}$$

The Q-function is tabulated in Appendix IV.

Both the Q-function and the error function contain the information necessary to evaluate integrals of Gaussian density functions. Some feel that the Q-function is more satisfying and easier to work with, since the integrand is a normalized Gaussian density.

Using the Q-function, let us now find the probability that a random variable is between two limits:

$$\Pr\{x_1 < X \le x_2\} = \frac{1}{\sqrt{2\pi}\sigma} \int_{x_1}^{x_2} \exp\left[\frac{(x-m)^2}{2\sigma^2}\right] dx \tag{4.40}$$

With the Q-function, the required change of variables yields

$$u = \frac{x - m}{\sigma}$$

$$\Pr\{x_1 < X \le x_2\} = \frac{1}{\sqrt{2\pi}} \int_{\frac{x_1 - m}{\sigma}}^{\frac{x_2 - m}{\sigma}} e^{-u^2/2}\, du \tag{4.41}$$

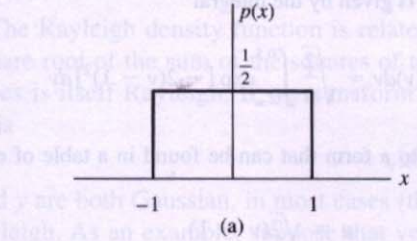
$$= Q\left(\frac{x_1 - m}{\sigma}\right) - Q\left(\frac{x_2 - m}{\sigma}\right)$$

Although the error function has traditionally been much more common than the Q-function, there are indications that the Q-function will predominate in the future. It makes absolutely no difference which one you use to solve a problem. (You'll get the same answer either way.) The only question you should have is which type of table is more readily available. Of course, if you use the wrong table, you will get the wrong answer.

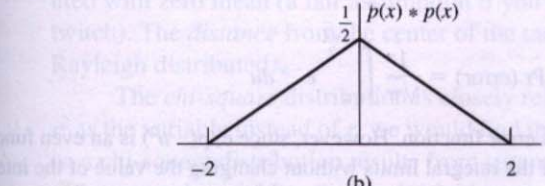
Now that we are familiar with the Gaussian density, let us return to the discussion of why it occurs so frequently in the real world. It results whenever a large number of factors contribute to an end result, as in the case of static in broadcast radio. Two conditions must be satisfied before the sum of many random variables starts to appear Gaussian. The first relates to the individual variances and to their infinite sum: The sum must approach infinity as the number of variables added together approaches infinity. The second condition is satisfied if the component densities go to zero outside some range. (This is a sufficient, but not a necessary, condition.) Since all quantities we deal with in the real world have bounded ranges, they satisfy the second condition.

Although we do not prove the central limit theorem here, one simple example is often given to indicate the reasonableness of the theorem. Suppose that we add together independent uniform random variables, where each is distributed between -1 and $+1$, as shown in Fig. 4.8(a). This is an example of *independent identically distributed* (*iid*) variables. When the first two variables are added, it can be shown that the resulting density is the convolution of the two original densities, as shown in Fig. 4.8(b). This doesn't yet look very much like the Gaussian curve, but we have summed only two variables.

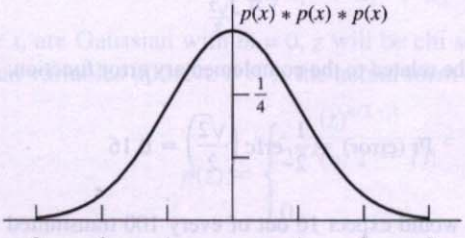
If a third variable is added, the ramp of Fig. 4.8(b) is convolved once more with the uniform density to get the parabolic curve of Fig. 4.8(c). At this point, the curve bears a re-



(a)



(b)



(c)

Figure 4.8 The central limit theorem.

semblance to the bell-shaped Gaussian density. As more and more variables are added, the agreement becomes closer and closer. In fact, the series converges quite rapidly to a Gaussian density, even if the densities we start with are not uniform.

Example 4.7

A binary communication system is a communication system that sends only one of two possible messages. A simple example of a binary system is one in which either *zero volts* or *one* volt is sent. Consider such a system in which the transmitted voltage is corrupted by additive atmospheric noise. If the receiver receives anything above $\frac{1}{2}$ volt (i.e., the midpoint), it assumes that a *one* was sent. If it receives anything below $\frac{1}{2}$ volt, it assumes that a *zero* was sent. Measurements show that if one volt is transmitted, the received signal level is random and has a Gaussian density with $m = 1$ and $\sigma = \frac{1}{2}$. Find the probability that a transmitted *one* will be interpreted as a *zero* at the receiver (i.e., find the probability of a *bit error*).

Solution: The received signal level has a Gaussian density with $m = 1$ and $\sigma^2 = (\frac{1}{2})^2$. Thus, if we designate the random variable as V, we have

$$P_V(v) = \sqrt{\frac{2}{\pi}} \exp\left[\frac{-(v-1)^2}{2(0.5)^2}\right]$$

Since any value received below a level of 0.5 is called 0, the probability that a transmitted 1 will be interpreted as a 0 at the receiver is simply the probability that the random variable V is less than 0.5. This is given by the integral

$$\int_{-\infty}^{0.5} P_V(v)dv = \sqrt{\frac{2}{\pi}} \int_{-\infty}^{0.5} \exp\left[-2(v-1)^2\right] dv$$

To reduce this equation to a form that can be found in a table of error functions, we make the change of variable

$$u = \sqrt{2}(v-1)$$

to get

$$\Pr(\text{error}) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{-\frac{\sqrt{2}}{2}} e^{-u^2} du$$

This is not yet in the form of an error function. However, since $\exp(-u^2)$ is an even function, we can take the mirror image of the integral limits without changing the value of the integral:

$$\Pr(\text{error}) = \frac{1}{\sqrt{\pi}} \int_{\frac{\sqrt{2}}{2}}^{\infty} e^{-u^2} du$$

This is now seen to be related to the complementary error function:

$$\Pr(\text{error}) = \frac{1}{2} \operatorname{erfc}\left(\frac{\sqrt{2}}{2}\right) = 0.16$$

Thus, on the average, one would expect 16 out of every 100 transmitted 1's to be misinterpreted as 0's at the receiver. This is an extremely poor level of performance.

4.2.2 Rayleigh Density Function

The *Rayleigh density function* is defined as

$$P_X(x) = \begin{cases} \dfrac{x}{K^2} \exp\left(\dfrac{-x^2}{2K^2}\right), & x > 0 \\ 0, & x < 0 \end{cases} \tag{4.42}$$

where K is a given constant. Figure 4.9 shows the density function for two different values of K.

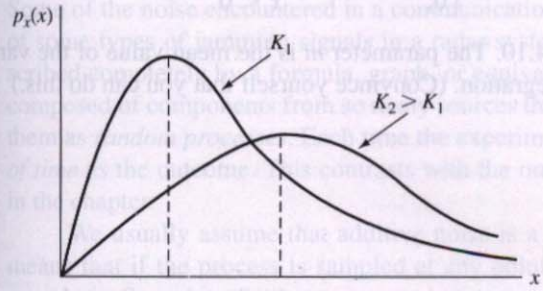


Figure 4.9 Rayleigh density function.

The Rayleigh density function is related to the Gaussian density function. In fact, the square root of the sum of the squares of two zero-mean Gaussian-distributed random variables is itself Rayleigh. If we transform from rectangular to polar coordinates, the radius is

$$r = \sqrt{x^2 + y^2}$$

If x and y are both Gaussian, in most cases (the restriction is one of *independence*) r will be Rayleigh. As an example, suppose that you were throwing darts at a target on a dart board and that the horizontal and vertical components of your error were Gaussian distributed with zero mean (a fair assumption if you are not biased by wind, gravity, or a muscle twitch). The *distance* from the center of the target to the position of the dart would then be Rayleigh distributed.

The *chi-square* distribution is closely related to the Rayleigh. If we were to consider r^2 as the variable instead of r, we would find that variable to be chi-square distributed. That is, a chi-square distribution results from summing the squares of Gaussian variables. If we sum two such variables, the result is chi square with *two degrees of freedom*. In general, if

$$z = x_1^2 + x_2^2 + x_3^2 + \ldots + x_n^2$$

and the x_i are Gaussian with $m = 0$, z will be chi square with n degrees of freedom. If the Gaussian variables all have $\sigma = 1$, the actual form of the density is given by

$$p(z) = \begin{cases} \dfrac{(z)^{n/2-1}}{2^{n/2}(n/2-1)!} e^{-z/2}, & z > 0 \\ 0, & z < 0 \end{cases} \tag{4.43}$$

The "!" in Eq. (4.43) indicates the *factorial* operation.

4.2.3 Exponential Random Variables

Occasionally, we deal with random variables that are exponentially distributed. This occurs in problems where we view a pattern of waiting times. (Such a pattern is important in communication network traffic studies.) It also shows up in examining the life of some systems, where we are interested in the *mean time between failures* (*MTBF*).

The exponential density is defined as

$$p_X(x) = \begin{cases} \frac{1}{m} e^{-x/m}, & x > 0 \\ 0, & x < 0 \end{cases} \tag{4.44}$$

This density is shown in Fig. 4.10. The parameter m is the mean value of the variable, as can be verified by a simple integration. (Convince yourself that you can do this.)

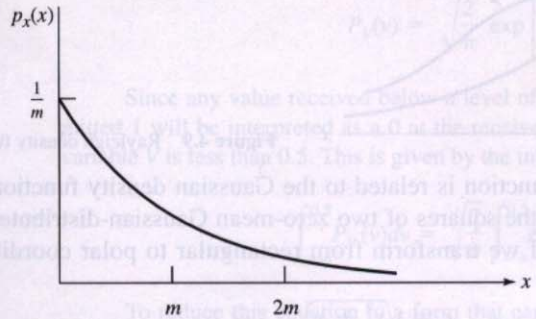


Figure 4.10 Exponential density function.

4.2.4 Ricean Density Function

In Section 4.2.2, we stated that the Rayleigh density function results when we take the square root of the sum of the squares of two zero-mean Gaussian densities. When we analyze digital communication systems in the absence of a signal (i.e., when noise alone is present), we often are dealing with variables that are Rayleigh distributed.

We also encounter situations where a transmitted signal is embedded in noise (which adds in the channel). In these cases, our receivers will sometimes effectively take the square root of the sum of the squares of two quantities. One of these will be zero-mean Guassian distributed, but the other results from an addition of signal and noise. The quantity we then observe is of the form

$$z = \sqrt{[s + x]^2 + y^2} \tag{4.45}$$

In this equation, x and y are zero-mean Gaussian random functions, and s is the signal. If s is zero, z follows a Rayleigh density function. If s is not zero, z follows a more complex density known as a *Ricean density*. This density is given by the equation

$$p(z) = \frac{z}{\sigma^2} \exp\left[-\frac{1}{2\sigma^2}(z^2 + s^2) \right] I_0\left(\frac{sz}{\sigma^2}\right) \tag{4.46}$$

In Eq. (4.46), s is the value of the signal (this will normally be a specific time sample $s(T)$), and I_0 is a *modified Bessel function of zero order*. We discuss Bessel functions in Chapter 6 (when we deal with FM). For now, you can think of them as functions you look up in a table. You simply plug in s, z, and σ^2 to find the argument of the Bessel function, and then you look in a table of Bessel functions. Note that when $s = 0$, Eq. (4.46) reduces to the Rayleigh density [$I_0(0) = 0$].

4.3 RANDOM PROCESSES

Some of the noise encountered in a communication system is *deterministic*, as in the case of some types of jamming signals in a radar system. In these cases, the noise can be described completely by a formula, graph, or equivalent technique. Other types of noise are composed of components from so many sources that we find it more convenient to analyze them as *random processes*. Each time the experiment is performed, we observe a *function of time* as the outcome. This contrasts with the one-dimensional outcomes studied earlier in the chapter.

We usually assume that additive noise is a random process that is *Gaussian*. This means that if the process is sampled at any point in time, the probability density of the sample is Gaussian. Such an assumption proves important, because if we start with a Gaussian random process and put it through any linear system, the system output will also be a Gaussian process. This fact will be used many times in examining the performance of various digital communication systems.

In addition to knowing that the noise process is Gaussian, it will be necessary to know something about the relationship between various time samples. We do this by examining the *autocorrelation function* and its Fourier transform, known as the *power spectral density*.

Until this point, we have considered single random variables. All averages and related parameters were simply numbers. Now we shall add another dimension to the study: time. Instead of talking about numbers only, we will now be able to characterize random functions. The advantages of such a capability should be clear. Our approach to random function analysis begins with the consideration of discrete-time functions, since these will prove to be a simple extension of random variables.

Imagine a single die being flipped 1,000 times. Let X_i be the random variable assigned to the outcome of the ith flip. Now list the 1,000 values of the random variables

$$X_1, X_2, X_3, \ldots, X_{999}, X_{1,000}$$

For example, if the random variable is assigned to be equal to the number of dots on the top face after flipping the die, a typical list might resemble the following:

$$4,6,3,5,1,4,2,5,3,1,4,5,\ldots.$$

Suppose that all possible sequences of random variables are now listed. We would then have a collection of $6^{1,000}$ entries, each one resembling the sequence shown. This collection is known as the *ensemble* of possible outcomes. Together with associated statistical properties, the ensemble forms a *random process*. In this particular example, the process is a discrete-valued and discrete-time process.

If we were to view one digit—say, the third entry—a random variable would result. In this example, the random variable would represent that assigned to the outcome of the third flip of the die.

We can completely describe the preceding random process by specifying the probability density function

$$p(x_1, x_2, x_3, \ldots, x_{999}, x_{1,000})$$

This 1,000-dimensional probability density function is used in the same way as a one-dimensional probability density function. Its integral must be unity:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} p(x_1, x_2, x_3, \ldots, x_{999}, x_{1,000})dx_1 dx_2 \ldots dx_{1,000} = 1 \tag{4.47}$$

The probability that the variables fall within any specified 1,000-dimensional volume is the integral of the probability density function over that volume. The expected value of any single variable can be found by integrating the product of that variable and the probability density function. Thus, the expected value of x_1 is

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} x_1 p(x_1, x_2, x_3, \ldots, x_{999}, x_{1,000})dx_1 dx_2 \ldots dx_{1,000} \tag{4.48}$$

This is known as a *first-order average*.

In a similar manner, the expected value of any multidimensional function of the variables is found by integrating the product of that function and the density function. Thus, the expected value of the product x_1x_2 is

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} x_1 x_2\, p(x_1, x_2, x_3, \ldots, x_{999}, x_{1,000})dx_1 dx_2 \ldots dx_{1,000} \tag{4.49}$$

This is a *second-order average*.

We can continue this process for higher order averages. In many instances, however, it proves sufficient to specify only the first- and second-order averages (moments). That is, one would specify

$$E\{x_i\} \quad \text{and} \quad E\{x_i x_j\} \qquad \text{for all } i \text{ and } j$$

Since we are interested in continuous functions of time, we now extend our example to the case of an infinite number of random variables in each list and an infinite number of lists in the ensemble. It may be helpful to refer back to the simple case (i.e., 1,000 flips of the die) from time to time.

The most general form of the random process results if our simple experiment yields an infinite range of values as the outcome and if the period between performances of the experiment approaches zero.

Another way to arrive at a random process is to perform a discrete experiment, but assign a *function of time* instead of a number to each outcome. When the samples of the process are functions of time, we call this a *stochastic process*. As an example, consider an experiment defined by picking a 6-volt dc generator from an infinite inventory in a warehouse. The voltage of the selected (call it the ith) generator, $v_i(t)$, is then displayed on an

oscilloscope. The waveform $v_i(t)$ is a *sample function* of the process. It will not be a perfect constant of 6 volts, due to imperfections in the generator construction and also due to rf pickup when the wires are acting as an antenna. Since we assume an unlimited inventory of generators, there is an infinite number of possible sample functions of the process. This infinite group of samples forms the ensemble. Each time we choose a generator and measure its voltage, a sample function from the infinite ensemble results.

If we were to sample the voltage at a specific time $t = t_0$, the sample $v(t_0)$ would be a random variable. Since $v(t)$ is assumed to be continuous with time, there is an infinity of random variables associated with the process.

Having introduced the concept with the generator example, let us now speak in general terms. Let $x(t)$ represent a stochastic process. $x(t)$ is then an infinite ensemble of all possible sample functions. For every specific value of time $t = t_0$, $x(t_0)$ is a random variable.

Suppose we now examine the first- and second-order averages of the process. Note that instead of having a discrete list of numbers, as in the die example, we have continuous functions of time. The first moment is then a function of time. This mean value is given by

$$m(t) = E[x(t)] \tag{4.50}$$

The second moments are found by averaging the product of two different time samples. We use the symbol R_x for this moment and call it the *autocorrelation*. The autocorrelation is then given by

$$R_x(t_1,t_2) = E[x(t_1)x(t_2)] \tag{4.51}$$

Both the mean and autocorrelation must be thought of as averages taken over the entire ensemble of functions of time. To find $m(t_0)$, we must average all samples across the ensemble at time t_0. To find the autocorrelation, we must average the product of $x(t_1)$ and $x(t_2)$ across the ensemble. This is generally difficult, and the ensemble averages are virtually impossible to ascertain experimentally.

In practice, we could find the mean by measuring the voltage of a great number of generators at time t_0 and average the resulting numbers. For the example of dc generators, we would expect this average to be independent of t_0. Indeed, most processes we consider have mean values that are independent of time.

A process with *overall statistics* that are independent of time is called a *stationary* (or *strict-sense stationary*) process. If only the mean and second moment are independent of time, the process is *wide-sense stationary*. Since we are interested primarily in power, and power depends upon the second moment, wide-sense stationarity will be sufficient for our analyses.

If a process $x(t)$ is stationary, then the shifted process $x(t - T)$ has the same statistics, independently of the value of T. Clearly, for a stationary process, $m(t_0)$ cannot depend upon t_0.

Viewing the autocorrelation of a stationary process, we have

$$R_x(t_1,t_2) = E\{x(t_1)x(t_2)\} = E\{x(t_1 - T)x(t_2 - T)\} \tag{4.52}$$

In the last equality, we have shifted the process by an amount T. If we now let $T = t_1$, we find that

$$R_x(t_1,t_2) = E\{x(0)x(t_2 - t_1)\} \tag{4.53}$$

Equation (4.53) indicates that the autocorrelation of a stationary (the wide sense is sufficient) process depends only on the time spacing t_2-t_1 between the two samples. That is, the left-hand time point can be placed anywhere, and as long as the right-hand point is separated from this by t_2-t_1, the autocorrelation remains unchanged. Since the independent variable of the autocorrelation is effectively one dimensional instead of two dimensional, we use the argument τ and refer to the autocorrelation of a stationary process as $R_x(\tau)$. Thus,

$$R_x(\tau) = E\{x(t)x(t-\tau)\} = E\{x(t)x(t+\tau)\} \tag{4.54}$$

The last equality results from adding τ to each of the arguments. This shows that autocorrelation is an even function.

If t_2 and t_1 are widely separated such that $x(t_1)$ and $x(t_2)$ are independent, the autocorrelation reduces to

$$R_x(\tau) = E\{x(t)x(t-\tau)\} = E\{x(t)\}\,E\{x(t-\tau)\} = m^2 \tag{4.55}$$

The average value of a random function of time is the dc value, and most communication channels will not pass dc. (They contain bandpass filters.) Therefore, most of the processes we consider have mean values equal to zero. In that case, the value of τ at which $R_x(\tau)$ goes to zero represents the time over which the process is correlated. If two samples are separated by this length of time, one sample has no effect upon the other.

Example 4.8

Suppose you are given a stochastic process $x(t)$ with mean value m and autocorrelation $R_x(\tau)$. This tells you immediately that the process is at least wide-sense stationary. If it were not, the mean and autocorrelation could not have been given in the form they were. Find the mean and autocorrelation of the process

$$y(t) = x(t) - x(t-T)$$

Solution: To solve problems of this type, we need only recall the definition of the mean and autocorrelation and the fact that taking expected values is a linear operation. We then have

$$\begin{aligned} m_y &= E\{y(t)\} = E\{x(t) - x(t-T)\} \\ &= E\{x(t)\} - E\{x(t-T)\} \\ &= m_x - m_x = 0 \\ R_y(\tau) &= E\{y(t)y(t+\tau)\} \\ &= E\{[x(t) - x(t-T)][x(t+\tau) - x(t+\tau-T)]\} \\ &= E\{x(t)x(t+\tau)\} - E\{x(t)x(t+\tau-T)\} \\ &\quad - E\{x(t-T)x(t+\tau)\} + E\{x(t-T)x(t+\tau-T)\} \\ &= R_x(\tau) - R_x(\tau-T) - R_x(\tau+T) + R_x(\tau) \\ &= 2R_x(\tau) - R_x(\tau-T) - R_x(\tau+T) \end{aligned}$$

We note that the process $y(t)$ is wide-sense stationary. If this were not the case, both m_y and R_y would be functions of t.

Example 4.9

Consider the experiment of starting a sinusoidal generator of deterministic frequency f_o and amplitude A. The exact starting time is random. Thus,

$$x(t) = A \sin (2\pi f_0 t + \theta)$$

where the phase θ is a random variable with uniform density, as shown in Fig. 4.11. Find the autocorrelation of the random process.

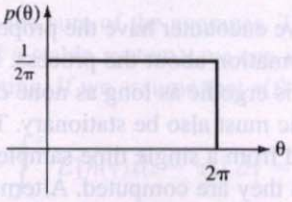


Figure 4.11 Density of phase for Example 4.9.

Solution: The autocorrelation is found directly from the definition:

$$R_x(\tau) = E\{x(t)x(t+\tau)\}$$

$$= E\{A^2 \sin (2\pi f_0 t + \theta) \sin [2\pi f_0(t+\tau) + \theta]\}$$

We can use trigonometric identities to rewrite this as

$$R_x(\tau) = E\left\{\frac{A^2}{2}(\cos 2\pi f_0 \tau - \cos [2\pi f_0(2t+\tau) + \theta])\right\}$$

The term $(A^2/2)(\cos 2\pi f_0\tau)$ is not random, so its expected value is itself. The expected value of $\cos[2\pi f_0 (2t + \tau) + \theta]$ is found by integrating its product with the density of θ. For the entire expression, we obtain

$$R_x(\tau) = \frac{1}{2}A^2 \cos 2\pi f_0\tau - \frac{A^2}{4\pi}\int_0^{2\pi} \cos [2\pi f_0(2t+\tau) + \theta]\, d\theta$$

The second term on the right represents the integral of a cosine function over two entire periods. This integral is equal to zero. Thus,

$$R_x(\tau) = \frac{1}{2}A^2 \cos 2\pi f_0\tau$$

Time Averages

Suppose you were asked to find the average value of the voltage in the dc generator example. You would have to measure the voltage of many generators (at any given time) and then compute the average. Once you were told that the process is stationary, you would probably be tempted to take one generator and average its voltage over a large time inter-

val. Using either technique, you would expect to get 6 volts as the result. That is, you would reason that

$$m_v = E\{v(t)\} = \lim_{T\to\infty} \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} v_i(t)\,dt \tag{4.56}$$

Here, $v_i(t)$ is one sample function of the ensemble. This approach does not always result in the correct answer. Suppose, for example, that one of the generators was burned out and you happened to choose that particular generator. You would erroneously think that $m_v = 0$.

Most of the processes we encounter have the property that any sample function contains all of the essential information about the process. Such processes are known as *ergodic*. The generator process is ergodic as long as none of the generators is "exceptional."

A process that is ergodic must also be stationary. This is so because, once we agree that all averages can be found from a single time sample, the averages can no longer be a function of the time at which they are computed. Alternatively, a stationary process need not be ergodic. (Consider the aforementioned burned-out generator.)

The autocorrelation of an ergodic process is given by

$$R_x(\tau) = E\{x(t)x(t+\tau)\} = \lim_{T\to\infty} \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} x(t)x(t+\tau)\,dt \tag{4.57}$$

The autocorrelation is a function of time. We define $G(f)$ as the Fourier transform of the autocorrelation. $G(f)$ is called the *power spectral density* for reasons that will become obvious in a moment. We have

$$G_x(f) = \mathcal{F}[R_x(t)] = \int_{-\infty}^{\infty} R_x(t)e^{-j2\pi ft}dt \tag{4.58}$$

The autocorrelation is then the inverse transform of the power spectral density:

$$R_x(t) = \mathcal{F}^{-1}[G_x(f)] = \int_{-\infty}^{\infty} G_x(f)\,e^{j2\pi ft}\,df = 2\int_0^{\infty} G_x(f)e^{j2\pi ft}\,df \tag{4.59}$$

In the last equality, we have doubled the positive half-range of the integral. This results from the fact that the power spectral density must be real and even, since the autocorrelation is even.

We can now relate the average power to the power spectral density:

$$P_{\text{av}} = E\{x^2(t)\} = R_x(0) = 2\int_0^{\infty} G_x(f)\,df \tag{4.60}$$

Equation (4.60) is a very important result. It says that to find the power of a random function of time, we integrate the power spectral density over all positive values of f (the frequency variable) and then double the result.

The other important result we need is the effect that a filter has on the power of a random signal. If a stochastic process forms the input to a filter, as shown in Fig. 4.12, the out-

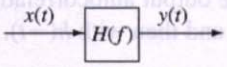


Figure 4.12 Stochastic process as input to filter.

put is also a stochastic process. That is, each sample function of the input process yields a sample function of the output process. We wish to find the statistics of the output process.

We begin with the mean value,

$$E\{y(t)\} = E\left\{\int_{-\infty}^{\infty} h(\tau)x(t-\tau)d\tau\right\} \tag{4.61}$$

The average of a sum is the sum of the averages. Therefore, with some broad restrictions (a finite mean value and a stable system), we can interchange the order of taking the expected value and integrating. If we assume that $x(t)$ is stationary, we have

$$E\{y(t)\} = \int_{-\infty}^{\infty} E\{h(\tau)x(t-\tau)\}\,d\tau = \int_{-\infty}^{\infty} h(\tau)E\{x(t-\tau)\}\,d\tau \tag{4.62}$$

$$= m_x\int_{-\infty}^{\infty} h(\tau)\,d\tau = m_x H(0)$$

Most of the random processes we encounter in this text have zero average value. If the input to the filter has zero average value, the output mean is also zero.

We now evaluate the autocorrelation of the output process. We assume that the input process is stationary. Then

$$R_y(t_1,t_2) = E\{y(t_1)y(t_2)\}$$
$$= E\left\{\int_{-\infty}^{\infty} h(\tau)x(t_1-\tau)d\tau \int_{-\infty}^{\infty} h(\tau)x(t_2-\tau)d\tau\right\} \tag{4.63}$$

Once again, we interchange the order of taking the expected value and integrating. We combine the two integrals (using two different symbols for the dummy variable of integration) to get

$$R_y(t_1,t_2) = \int_{-\infty}^{\infty} h(\tau_1)d\tau_1 \int_{-\infty}^{\infty} E\{x(t_1-\tau_1)x(t_2-\tau_2)\}h(\tau_2)d\tau_2$$
$$= \int_{-\infty}^{\infty} h(\tau_1)d\tau_1 \int_{-\infty}^{\infty} R_x(t_1-t_2+\tau_2-\tau_1)h(\tau_2)d\tau_2 \tag{4.64}$$

Note that the result depends, not on the values of t_1 and t_2, but only on their difference. Therefore, the output process is wide-sense stationary. The autocorrelation is a function of only one variable, the spacing between the two time points. We have used the notation τ for this spacing. Doing the same here, we find that the result becomes

$$R_y(\tau) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} R_x(\tau-\tau_1+\tau_2)h(\tau_1)h(\tau_2)d\tau_1 d\tau_2 \tag{4.65}$$

Equation (4.65) shows that the output autocorrelation is the result of convolving the input autocorrelation, first with $h(t)$ and then with $h(-t)$. Therefore,

$$R_y(t) = R_x(t)*h(t)*h(-t) \tag{4.66}$$

Taking the Fourier transform of this equation, and recognizing that the Fourier transform of $h(-t)$ is the complex conjugate of the transform of $h(t)$, we have

$$G_y(f) = G_x(f)H(f)H^*(f) = G_x(f)|H(f)|^2 \tag{4.67}$$

Equation (4.67) has the intuitive interpretation. $G(f)$ is the power spectral density. It is not surprising that the output power spectral density is weighted by the square of the magnitude of the transfer function, since this is what would happen to the power of a single sinusoid that goes through the filter.

Example 4.10

A signal that is received is made up of two components: signal and noise. That is,

$$r(t) = s(t) + n(t)$$

The signal can be considered a sample of a random process because random amplitude fluctuations are introduced by turbulence in the air. You are told that the autocorrelation of the signal process is

$$R_s(\tau) = 2e^{-|\tau|}$$

The noise is a sample function of a random process with autocorrelation

$$R_n(\tau) = e^{-2|\tau|}$$

Both processes have zero mean value, and they are independent of each other.

Find the autocorrelation and total power of $r(t)$.

Solution: From the definition of autocorrelation, we have

$$\begin{aligned} R_r(\tau) &= E\{r(t)r(t+\tau)\} \\ &= E\{[s(t) + n(t)][s(t+\tau)n(t+\tau)]\} \\ &= E\{s(t)s(t+\tau)\} + E\{s(t)n(t+\tau)\} \\ &= E\{n(t)s(t+\tau)\} + E\{n(t)n(t+\tau)\} \end{aligned}$$

Since the signal and noise are independent,

$$E\{s(t+\tau)n(t)\} = E\{s(t+\tau)\}E\{n(t)\} = 0$$

and

$$E\{s(t)n(t+\tau)\} = E\{s(t)\}E\{n(t+\tau)\} = 0$$

Finally, the autocorrelation is

$$R_r(\tau) = R_s(\tau) + R_n(\tau) = 2e^{-|\tau|} + e^{-2|\tau|}$$

The total power of $r(t)$ is $R_r(0)$, or 3 watts.

Example 4.11 (the random telegraph signal)

Evaluate the autocorrelation of the *random telegraph waveform*, as shown in Fig. 4.13. This is a binary waveform that can take on one of two values, $+A$ or $-A$. The probabilities of each

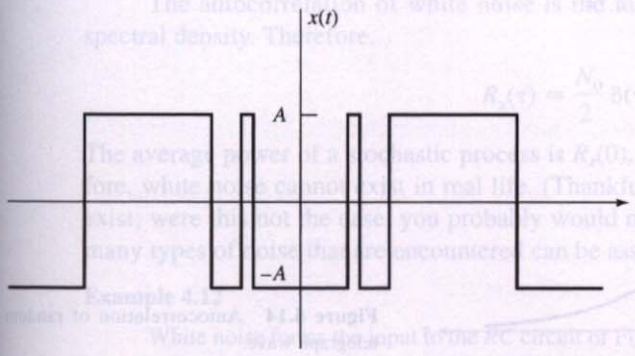


Figure 4.13 Random telegraph waveform.

of these values are equal (i.e., $\frac{1}{2}$). Assume that transitions occur randomly and that there is an average of λ transitions per second. The probability that n transitions occur in a positive time interval τ is given by a *Poisson distribution*,

$$\Pr(n,\tau) = \frac{(\lambda\tau)^n}{n!} e^{-\lambda\tau}$$

Solution: We first find the autocorrelation of the process:

$$R_x(\tau) = E\{x(t)x(t+\tau)\}$$

The product inside the braces is either $+A^2$ or $-A^2$. The plus sign obtains if there is an even number of transitions in the interval, and the minus sign obtains if there is an odd number. For a given value of τ, the probability of an even number of transitions is found by summing the Poisson probability distribution over all even values of n. Similarly, the probability of an odd number of transitions is the sum of the Poisson distribution over all odd values of n. Therefore,

$$\text{PR(even)} = e^{-\lambda\tau} \sum_{\substack{n=0 \\ n=\text{even}}}^{\infty} \frac{(\lambda\tau)^n}{n!}$$

$$\text{PR(odd)} = e^{-\lambda\tau} \sum_{\substack{n=1 \\ n=\text{odd}}}^{\infty} \frac{(\lambda\tau)^n}{n!}$$

The autocorrelation is then

$$\begin{aligned} R_x(\tau) &= A^2 \Pr(\text{even}) - A^2 \Pr(\text{odd}) \\ &= A^2 e^{-\lambda\tau} \sum_{n=0}^{\infty} (-1)^n \frac{(\lambda\tau)^n}{n!} \\ &= A^2 e^{-\lambda\tau} e^{-\lambda\tau} = A^2 e^{-2\lambda\tau} \end{aligned}$$

The result applies for positive time invervals. We know that the autocorrelation must be an even function, so we can write the autocorrelation as

$$R_x(\tau) = A^2 e^{-2\lambda|\tau|}$$

This result is shown in Fig. 4.14.

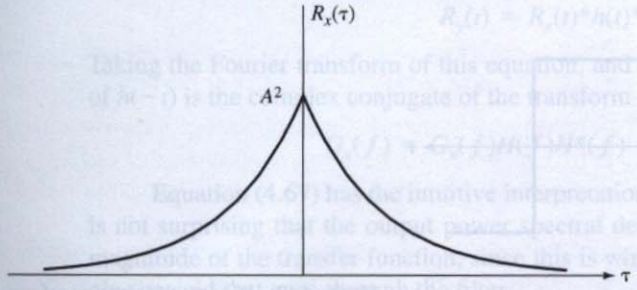


Figure 4.14 Autocorrelation of random telegraph wave.

4.4 WHITE NOISE

Suppose that $x(t)$ is a stochastic process with a constant power spectral density, as shown in Fig. 4.15. This process contains all frequencies "to an equal degree." Since white light is composed of all frequencies (colors), the process is known as *white noise*.

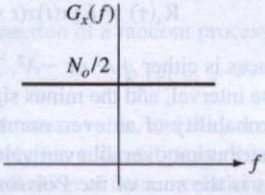


Figure 4.15 Power spectral density of white noise.

Suppose now that white noise forms the input to an ideal bandpass filter with a passband extending from a low-frequency cutoff of f_L to a high-frequency cutoff of f_H. Then the power of the output for the filter is (refer to Eq. (4.67))

$$P_{\text{out}} = 2\int_0^{\infty} G_x(f)|H(f)|^2\,df = 2\int_{f_L}^{f_H} G_x(f)\,df = 2\int_{f_L}^{f_H} \frac{N_0}{2}\,df = N_0(f_H - f_L) \qquad (4.68)$$

The output of the bandpass filter consists of all components of the input lying within the passband of the filter. The output power can therefore be considered to be that portion of the input power in the frequency range between f_L and f_H. We see from Eq. (4.68) that this is proportional to the bandwidth, with the proportionality factor being N_0. Therefore, N_0 is the *power per Hz* of the noise waveform. The total power in a band of frequencies is the product of N_0 with the bandwidth.[4]

[4]The power spectral density of Fig. 4.15 is known as the *two-sided power spectral density*. Since the power spectral density is real and even, and the equation for power contains a factor of two, we sometimes define a *one-sided power spectral density* with a value twice that of the two-sided density. The one-sided density of white noise therefore has a height of N_0 instead of $N_0/2$, and to find the power in a band of frequencies, we simply integrate (without the factor of two).

The autocorrelation of white noise is the inverse Fourier transform of the power spectral density. Therefore,

$$R_x(\tau) = \frac{N_0}{2}\delta(\tau) \tag{4.69}$$

The average power of a stochastic process is $R_x(0)$, which, for this case, is infinity. Therefore, white noise cannot exist in real life. (Thankfully, signals with infinite power do not exist; were this not the case, you probably would not be here to read this text.) However, many types of noise that are encountered can be assumed to be approximately white.

Example 4.12

White noise forms the input to the *RC* circuit of Fig. 4.16. Find the autocorrelation and power spectral density at the output of the filter.

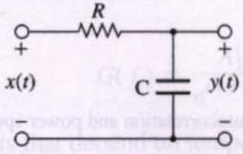


Figure 4.16 Circuit for Example 4.12.

Solution: The output power spectral density is the input density multiplied by the square of the magnitude of the transfer function:

$$G_y(f) = G_x(f)|H(f)|^2 = \frac{N_0/2}{1 + (2\pi f)^2 C^2 R^2}$$

The output autocorrelation is the inverse Fourier transform of $G_y(f)$. Therefore,

$$R_y(\tau) = \frac{N_0/2}{2RC}\exp\left(\frac{-|\tau|}{RC}\right)$$

Example 4.13

Repeat Example 4.12 for an ideal lowpass filter with cutoff frequency f_m, as shown in Fig. 4.17.

Solution: Once again, the output power spectral density is found from the input density by

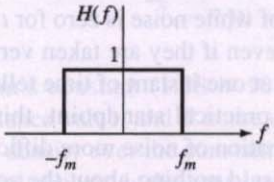


Figure 4.17 Ideal lowpass filter for Example 4.13.

multiplying it by the square of the magnitude of the transfer function:

$$G_y(f) = G_x(f)|H(f)|^2$$
$$= \begin{cases} N_0/2, & |f| < f_m \\ 0, & \text{otherwise} \end{cases}$$

The autocorrelation is the inverse Fourier transform of $G_y(f)$. Thus,

$$R_y(\tau) = \frac{N_0}{2} \frac{\sin 2\pi f_m \tau}{\pi \tau}$$

The output power spectral density and autocorrelation are shown in Fig. 4.18.

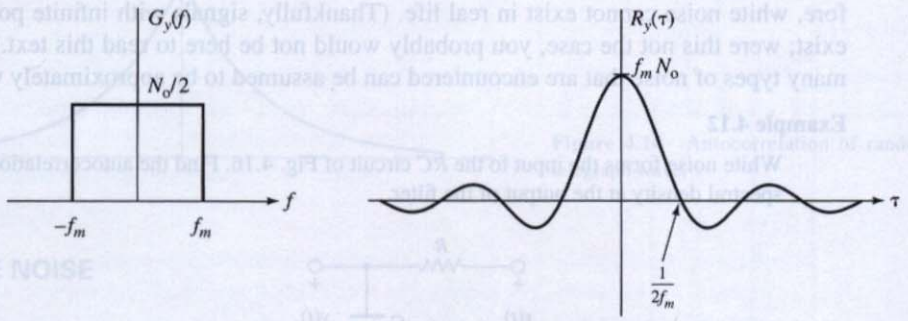


Figure 4.18 Autocorrelation and power spectrum for Example 4.13.

Suppose that in Example 4.13 the input process had a power spectral density as shown in Fig. 4.19, where $f_1 > f_m$. The output process would then be identical to that found in the example. Therefore, if a system exhibits an upper cutoff frequency, and the input noise has a flat spectrum up to this cutoff frequency, we can consider the input noise to be white. This will simplify the analysis.

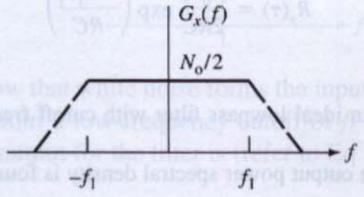


Figure 4.19 Nonwhite power spectral density.

Since the autocorrelation of white noise is zero for $t \neq 0$, two different time samples of white noise are uncorrelated even if they are taken very close together. Thus, knowing the sample value of white noise at one instant of time tells us absolutely nothing about its value an instant later.[5] From a practical standpoint, this is an unfortunate situation. It would appear to make the elimination of noise more difficult.

Up to this point, we have said nothing about the actual probability distributions of the process. We have talked only about the first and second moment. There is an infinity of random processes with the same first and second moments. (The analogy to mechanics is that, given the center of gravity and moment of inertia of an object, the exact shape can be any of an infinity of possibilities.) Each random variable $x(t_0)$ has a certain probability density. By considering only the mean and second moment, we are not telling the whole story.

[5]The zero correlation is a necessary, but not sufficient, condition for independence. However, for Gaussian processes, two uncorrelated samples are always independent.

Because of the central limit theorem, most processes we encounter are Gaussian. Once we know that a random variable is Gaussian, the density function is completely specified by its mean and second moment.

We now examine several types of noise encountered in communication systems and determine whether the white noise model is appropriate for these noise sources.

Thermal Noise

Thermal noise is produced by the random motion of electrons in a medium. The intensity of this motion increases with increasing temperature and is zero only at a temperature of absolute zero.

If the voltage across a resistor is examined using a sensitive oscilloscope, a random pattern will be displayed on the screen. The power spectral density of this random process is of the form

$$G(f) = \frac{A|f|}{e^{B|f|} - 1} \tag{4.70}$$

where A and B are constants that depend on temperature and other physical constants. Figure 4.20 shows the curve of Eq. (4.70). For frequencies below the knee of the curve, $G(f)$ is almost constant. If we operate in this frequency range, we can consider thermal noise to be white noise. Thermal noise appears to be approximately white up to extremely high frequencies, on the order of 10^{13} Hz. For frequencies within this range, the mean square value of the voltage across the resistor $[R(0)]$ has been shown to be

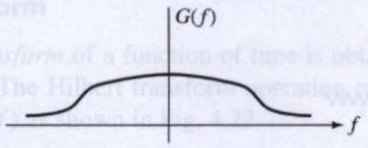


Figure 4.20 Power spectral density of thermal noise.

$$\overline{v^2} = R(0) = 4kTRB \tag{4.71}$$

where k is Boltzmann's constant (1.38×10^{-23} J/°K), T is the temperature in degrees Kelvin, R is the resistance value, and B is the observation bandwidth. This means that the height of the spectral density over the constant region is $2kTR$.

Of more practical concern is the actual power generated by a resistor. That is, if a resistor is connected to additional circuitry, how much noise power is generated in that additional circuitry? We know from basic circuit theory that this depends on the impedance of the external circuit. Specifically, the power transferred is a maximum when the load impedance matches the generator impedance. This yields the *maximum available power*, which (using a voltage divider relationship) is

$$N = \frac{v^2}{4R} = kTB$$

with a corresponding power spectral density of

$$G_n(f) = \frac{kT}{2} \tag{4.72}$$

Equation (4.72) yields the power spectral density of the available noise power from a resistor.

If we have a system with a number of noise-generating devices within it, we often refer to the system *noise temperature*, T_e, in degrees Kelvin. This is the temperature of a single noise source that would produce the same total noise power at the output.

If the input to the system contains noise, the system then adds its own noise to produce a larger output noise. The system *noise figure* is the ratio of noise power at the output to that at the input. It is usually expressed in dB. For example, a noise figure of 3 dB indicates that the system is adding an amount of noise equal to that which appears at the input, so the output noise power is twice that of the input.

Other Forms of Noise

Shot noise (or *quantum noise*) occurs because, although we think of current as being continuous, it is actually a discrete phenomenon. In fact, current occurs in discrete pulses each time an electron moves across an observation point. A plot of current as a function of time would resemble that of Fig. 4.21. Shot noise is the variation of current around the average value. As in the case of thermal noise, the power spectral density of shot noise is approximately flat within the range of frequencies of interest to us.

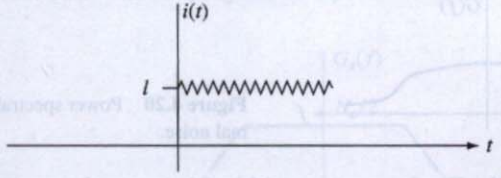


Figure 4.21 Shot noise.

Flicker noise (or *1/f noise*) occurs in electronic devices. It arises out of surface imperfections resulting from the fabrication process. Its power spectral density decreases inversely with increasing frequency. Flicker noise is most important at low frequencies (below about 100 Hz). Thus, although it cannot be approximated as white noise, it usually is negligible at the frequencies at which communication systems operate.

4.5 NARROWBAND NOISE

Most communication systems with which we deal contain bandpass filters. Therefore, white noise appearing at the input to the system will be shaped into bandlimited noise by the filtering operation. If the bandwidth of the noise is relatively small compared to the center frequency, we refer to this as *narrowband noise*. We have no problem deriving the power spectral density and autocorrelation of this noise, and these quantities are sufficient to analyze the effect of linear systems. However, we will often be dealing with multipliers, and the frequency analysis approach is not sufficient, since nonlinear operations are pres-

ent. In such cases, it proves useful to have a trigonometric expansion for the noise signals. The form of this expansion is

$$n(t) = x(t)\cos 2\pi f_0 t - y(t)\sin 2\pi f_0 t \tag{4.73}$$

In equation (4.73), $n(t)$ is the noise waveform and f_0 is a frequency (often the center frequency) within the band occupied by the noise. Since the sine and cosine vary by 90 degrees, $x(t)$ and $y(t)$ are known as the *quadrature components* of the noise.

Equation (4.73) can be derived by starting with exponential notation. We have

$$n(t) = \text{Re}\{r(t)e^{j2\pi f_0 t}\} \tag{4.74}$$

where $r(t)$ is a complex function with a low-frequency bandlimited Fourier transform, Re is the real part of the expression in brackets that follows it, and the exponential function has the effect of shifting the frequencies of $r(t)$ by f_0. Expanding the exponential by means of Euler's identity and letting $x(t)$ be the real part of $r(t)$ and $y(t)$ be the imaginary part, we have

$$\begin{aligned} n(t) &= \text{Re}\{[x(t) + jy(t)]\,(\cos 2\pi f_0 t + j\sin 2\pi f_0 t)\} \\ &= x(t)\cos 2\pi f_0 t - y(t)\sin 2\pi f_0 t \end{aligned} \tag{4.75}$$

This is the same as Eq. (4.73).

Solving Eq. (4.73) explicitly for $x(t)$ and $y(t)$ is not simple. One way to do so is by using Hilbert transforms.

Hilbert Transform

The *Hilbert transform* of a function of time is obtained by shifting all frequency components by $-90°$. The Hilbert transform operation can therefore be represented by a linear system, with $H(f)$ as shown in Fig. 4.22.

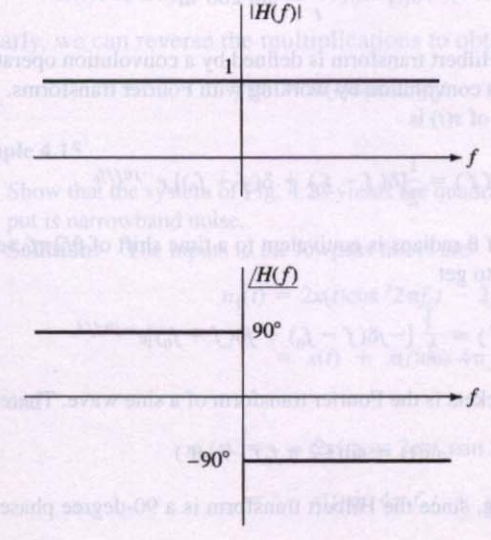


Figure 4.22 Hilbert transform operation.

Note that the phase function of a real system must be odd. The system function is then given by

$$H(f) = -j\,\mathrm{sgn}\,(f) \tag{4.76}$$

The impulse response of this system is the inverse transform of $H(f)$. This is given by

$$h(t) = \frac{1}{\pi t} \tag{4.77}$$

The Hilbert transform of $s(t)$ is then given by the convolution of $s(t)$ with $h(t)$. Let us denote the transform by $\hat{s}(t)$. Then

$$\hat{s}(t) = -\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{s(\tau)}{t-\tau}\,d\tau \tag{4.78}$$

If we take the Hilbert transform of a Hilbert transform, the effect in the frequency domain is to multiply the transform of the signal by $H^2(f)$. But $H^2(f) = -1$, so we return to the original signal which is a change of sign. This indicates that the inverse Hilbert transform equation is the same as the transform relationship, except with a minus sign. Therefore,

$$s(t) = -\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{\hat{s}(\tau)}{t-\tau}\,d\tau \tag{4.79}$$

Example 4.14

Find the Hilbert transform of the following time signals:

(a) $s(t) = \cos(2\pi f_0 t + \theta)$

(b)

$$s(t) = \frac{\sin 2\pi t}{t}\cos 200\,\pi t$$

(c)

$$s(t) = \frac{\sin 2\pi t}{t}\sin 200\,\pi t$$

Solution: Although the Hilbert transform is defined by a convolution operation, it is almost always easier to avoid time convolution by working with Fourier transforms.

(a) The Fourier transform of $s(t)$ is

$$S(f) = \frac{1}{2}[\delta(f-f_0) + \delta(f+f_0)]\,e^{-j\theta f/f_0}$$

Note that the phase shift of θ radians is equivalent to a time shift of $\theta/2\pi f_0$ seconds. We now multiply this by $-j\mathrm{sgn}(f)$ to get

$$\hat{S}(f) = \frac{1}{2}[-j\delta(f-f_0) + j\delta(f+f_0)]e^{-j\theta f/f_0}$$

The quantity in square brackets is the Fourier transform of a sine wave. Therefore,

$$\hat{s}(t) = \sin(2\pi f_0 t + \theta)$$

This result is not surprising, since the Hilbert transform is a 90-degree phase-shifting operation.

(b) Let

$$x(t) = \frac{\sin 2\pi t}{t}$$

The Fourier transform of $s(t)$ is then

$$S(f) = \frac{1}{2}X(f - 100) + \frac{1}{2}X(f + 100)$$

Since $X(f)$ is bandlimited to $f = \pm 1$, the first term in $S(f)$ occupies frequencies between 99 and 101 Hz, while the second term occupies frequencies between -101 and -99 Hz. When $S(f)$ is multiplied by $-j\,\text{sgn}(f)$, we find that

$$\hat{S}(f) = -\frac{1}{2}jX(f - 100) + \frac{1}{2}jX(f + 100)$$

The inverse transform yields

$$\hat{s}(t) = x(t)\sin 200\,\pi t = \frac{\sin 2\pi t}{t}\sin 200\,\pi t$$

(c) We use the fact that the Hilbert transform of a Hilbert transform is the negative of the original function. Therefore, by inspection, we have

$$\hat{s}(t) = -x(t)\cos 200\pi t = -\frac{\sin 2\pi t}{t}\cos 200\,\pi t$$

We are now ready to return to the solution of Eq. (4.73). If $x(t)$ and $y(t)$ are assumed to be bandlimited to frequencies below f_0, we can take the Hilbert transform of both sides of that equation to get

$$\hat{n}(t) = x(t)\sin 2\pi f_0 t + y(t)\cos 2\pi f_0 t \tag{4.80}$$

If Eq. (4.73) is multiplied by $\cos 2\pi f_0 t$ and Eq. (4.80) is multiplied by $\sin 2\pi f_0 t$, then when the two expressions are added together, $y(t)$ is eliminated, yielding

$$n(t)\cos 2\pi f_0 t + \hat{n}(t)\sin 2\pi f_0 t = x(t)[\cos^2 2\pi f_0 t + \sin^2 2\pi f_0 t] = x(t) \tag{4.81}$$

Similarly, we can reverse the multiplications to obtain

$$y(t) = \hat{n}(t)\cos 2\pi f_0 t - n(t)\sin 2\pi f_0 t \tag{4.82}$$

Example 4.15

Show that the system of Fig. 4.23 yields the quadrature components at the output when the input is narrowband noise.

Solution: The inputs to the lowpass filters are

$$n_1(t) = 2x(t)\cos^2 2\pi f_0 t - 2y(t)\sin 2\pi f_0 t\cos 2\pi f_0 t$$

$$= x(t) + x(t)\cos 4\pi f_0 t - y(t)\sin 4\pi f_0 t$$

and

$$n_2(t) = -2x(t)\cos 2\pi f_0 t\sin 2\pi f_0 t + 2y(t)\sin^2 2\pi f_0 t$$

$$= -x(t)\sin 4\pi f_0 t + y(t) - y(t)\cos 4\pi f_0 t$$

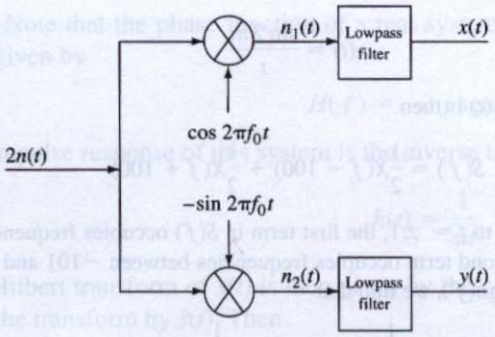


Figure 4.23 System to generate quadrature components.

The modulation theorem indicates that the Fourier transform of $x(t)\cos 4\pi f_0 t$ occupies a range around a frequency of $2f_0$. This is also true of the other terms with $4\pi f_0$ in the argument of the sinusoid. The lowpass filter is designed to pass the frequencies of $x(t)$ and $y(t)$, so it will reject these high-frequency terms. The outputs are therefore as shown on the diagram.

The autocorrelation of $x(t)$ and $y(t)$ can now be derived from Eqs. (4.81) and (4.82):

$$R_x(\tau) = R_y(\tau) = R_n(\tau)\cos 2\pi f_0\tau + \left(R_n(\tau) * \frac{1}{\pi\tau}\right)\sin 2\pi f_0\tau \tag{4.83}$$

Finally, we apply the modulation theorem to Eq. (4.83) to get

$$G_x(f) = G_y(f) = G_n(f - f_0) + G_n(f + f_0) \quad \text{for } f_0 - f_m < |f| < f_0 + f_m \tag{4.84}$$

Equation (4.84) is the key result that will enable us to calculate the effects of noise on AM and FM communication systems.

Example 4.16

Express the three narrowband noise processes of Fig. 4.24 in quadrature form, using f_0 as the center frequency.

Solution: We use Eq. (4.84) to immediately sketch the power spectral densities of $x(t)$ and $y(t)$. These are shown in Fig. 4.25. The noise is then expressed as

$$n(t) = x(t)\cos 2\pi f_0 t - y(t)\sin 2\pi f_0 t$$

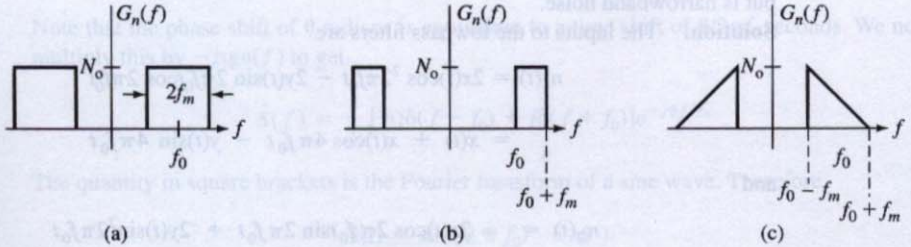


Figure 4.24 Noise processes for Example 4.16.

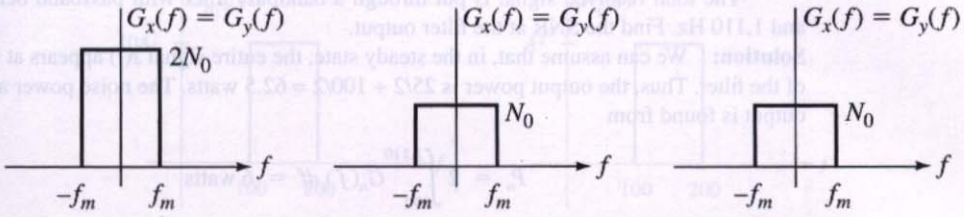


Figure 4.25 Power spectral density of quadrature components for Example 4.16.

4.6 SIGNAL-TO-NOISE RATIO

In many communication studies, the probabilistic parameter of interest is the *average power*. By itself, this quantity would not tell very much, since we could always modify the average power of a signal by putting it through an amplifier or an attenuator. A problem arises, however, because the received waveform usually consists of a desired signal plus noise. If we amplify or attenuate the total received waveform, we do the same thing to the noise as we do to the signal. The parameter of interest then is not the signal power, but the ratio of that power to the power of the unwanted noise. This is the *signal-to-noise ratio,* abbreviated as S/N or SNR.

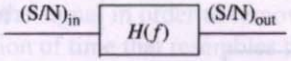


Figure 4.26 Signal-to-noise ratio.

Figure 4.26 shows a block diagram of a system with the input S/N and output S/N indicated. The ratio of these two SNRs gives some measure of the effectiveness of the system. We designate the ratio as ΔSNR, the *signal-to-noise improvement* of the system:

$$\Delta\text{SNR} = \frac{(S/N)_{out}}{(S/N)_{in}} \tag{4.85}$$

ΔSNR is often expressed in *decibels*, or dB, as

$$\Delta\text{SNR}_{dB} = 10 \log_{10}(\Delta\text{SNR}) \tag{4.86}$$

Example 4.17

A signal is given by

$$r(t) = s(t) + n(t)$$

where

$$s(t) = 5 \cos 2\pi \times 1{,}000t + 10 \cos 2\pi \times 1{,}100t$$

The noise $n(t)$ is white with power $N_0 = 0.05$ watt/Hz.

The total received signal is put through a bandpass filter with passband between 990 and 1,110 Hz. Find the SNR at the filter output.

Solution: We can assume that, in the steady state, the entire signal $s(t)$ appears at the output of the filter. Thus, the output power is 25/2 + 100/2 = 62.5 watts. The noise power at the filter output is found from

$$P_n = 2\int_{990}^{1,110} G_n(f)\,df = 6 \text{ watts}$$

The SNR is then

$$S/N = 62.5/6 = 10.4 = 10.17 \text{ dB}$$

Example 4.18

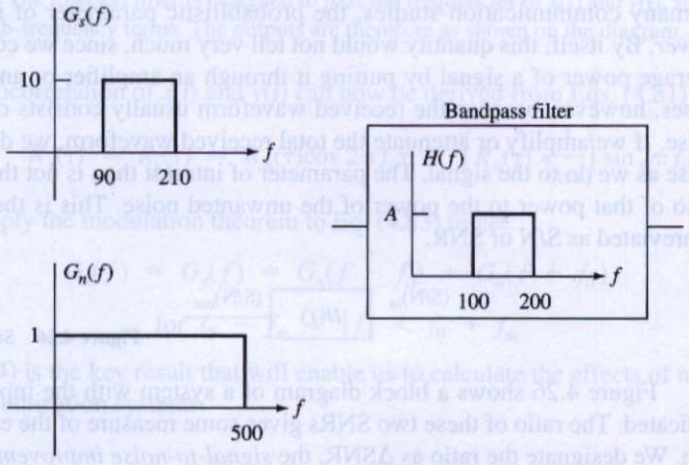


Figure 4.27 Bandpass filter and signals for Example 4.18.

The input to the bandpass filter shown in Fig. 4.27 is the sum of a signal and noise. The power spectral density of the signal and of the noise are as shown. Find the S/N improvement of the filter.

Solution: The input powers are found by integrating the corresponding densities:

Signal power in = 2,400 watts
Noise power in = 1,000 watts
S/N in = 2.4 or 4.8 dB

The output power spectral densities, which are illustrated in Fig. 4.28, result from multiplying the input densities by the square of the magnitude of the filter transfer function. Integration of these densities results in the following output powers:

Signal power out = $2{,}000A^2$ watts
Noise power out = $200A^2$ watts
S/N out = 10 or 10 dB

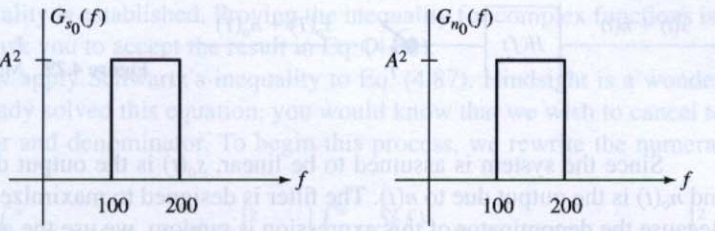


Figure 4.28 Output power spectral densities for Example 4.18.

The S/N improvement is

$$\Delta\text{SNR} = 10/2.4 = 4.17 \text{ or } 6.2 \text{ dB}$$

Note that we could have found the S/N improvement by subtracting the input S/N from the output S/N when both are expressed in dB. Note further that the answer is independent of A. This is true because both the noise and the signal are multiplied by A, so the ratio is unaffected.

4.7 MATCHED FILTER

There are a variety of operations we may wish to perform on a received signal. In some cases, we wish to filter the signal in order to remove as much noise as possible and therefore be left with a function of time that resembles the desired signal as closely as possible. In other situations, we may wish to maximize the output SNR without regard to preserving the shape of the signal waveform. That is, we may use a filter that significantly alters the shape of both the signal and the noise in a way that increases the SNR. Such a filter distorts the signal.

Analog receivers typically try to reconstruct the waveform as closely as possible, while digital receivers attempt to "pull" the signal out of background noise without regard to distortion.

In order to motivate this study, let us get way ahead of the game. Suppose you wish to send a list of 1's and 0's by speaking the words *one* and *none* into a microphone. Then, to send 1010, you would speak *one-none-one-none*. Noise adds to the transmitted signal, so let us assume that the receiver has difficulty in distinguishing between the two different words sent. Now suppose that you filter the received signal plus noise in a manner that blocks a good portion of the noise. But in the process, you change the transmitted *one* signal into a waveform that, when placed into a speaker, generates the word *start*. The same filter changes *none* to *halt*. Therefore, instead of hearing a highly noise-corrupted sequence consisting of repetitions of the words *one* and *none*, you hear a relatively uncorrupted signal consisting of *start-halt-start-halt*. You could still recover the original binary sequence in spite of what would be considered severe distortion of the signal waveform.

The *matched filter* is a linear system that maximizes the output SNR. We designate the input to the filter as $s(t) + n(t)$, and the resulting output is $s_o(t) + n_o(t)$. This is shown in Fig. 4.29.

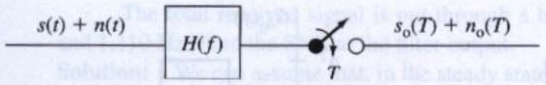


Figure 4.29 Matched filter.

Since the system is assumed to be linear, $s_o(t)$ is the output due to an input of $s(t)$, and $n_o(t)$ is the output due to $n(t)$. The filter is designed to maximize the ratio $s^2_o(T)/n^2_o(T)$. Because the denominator of this expression is random, we use the average value. The output SNR is then

$$\rho = \frac{s_0^2(T)}{n_0^2(T)} = \frac{\left| \int_{-\infty}^{\infty} S(f)H(f)e^{j2\pi fT}\,df \right|^2}{\int_{-\infty}^{\infty} |H(f)|^2 G_n(f)\,df} \tag{4.87}$$

The numerator of Eq. (4.87) is the square of the inverse Fourier transform of the product of the input transform with the system function. Thus, it is the square of the deterministic time sample $s_o(T)$. The $G_n(f)$ in the denominator is the power spectral density of the input noise. Thus, the denominator integrand is the power spectral density of the output noise. Note that we are integrating from $-\infty$ to $+\infty$ instead of doubling the integral from zero to ∞. We do this for reasons that will soon become clear.

We wish to choose $H(f)$ so as to maximize Eq. (4.87). This is a difficult maximization problem. (You cannot simply take the derivative and set it to zero, since we are trying to find a function rather than a value.) The choice is simplified if we apply *Schwartz's inequality* to the numerator. In doing so, we will be able to solve for $H(f)$ almost by inspection.

Schwartz's inequality states that for all functions $f(x)$ and $g(x)$,

$$\left| \int f(x)g(x)\,dx \right|^2 \leq \int |f^2(x)|\,dx \int |g^2(x)|\,dx \tag{4.88}$$

We will derive Schwartz's inequality for the special case of real functions by starting with the observation that

$$\int [f(x) - Tg(x)]^2\,dx \geq 0$$

for all real $f(x)$, $g(x)$, and T (i.e., we are integrating a non-negative function). Expanding this expression, we obtain

$$T^2 \int g^2(x)dx - 2T \int f(x)g(x)\,dx + \int f^2(x)\,dx \geq 0 \tag{4.89}$$

The left side of Eq. (4.89) is quadratic in T. Since the value can never be negative, the quadratic cannot have distinct real roots. Therefore, the discriminant cannot be positive. Thus,

$$4\left[\int f(x)g(x)\,dx\right]^2 - 4\int f^2(x)\,dx \int g^2(x)\,dx \leq 0 \tag{4.90}$$

and the inequality is established. Proving the inequality for complex functions is more difficult, so we ask you to accept the result in Eq. (4.88).

We now apply Schwartz's inequality to Eq. (4.87). Hindsight is a wonderful thing: Had you already solved this equation, you would know that we wish to cancel terms from the numerator and denominator. To begin this process, we rewrite the numerator of Eq. (4.87) as

$$\left|\int_{-\infty}^{\infty} S(f)H(f)e^{j2\pi fT}\,df\right|^2 = \left|\int_{-\infty}^{\infty} \frac{S(f)}{\sqrt{G_n(f)}} H(f)\sqrt{G_n(f)}e^{j2\pi fT}\,df\right|^2 \tag{4.91}$$

The square root operation is unambiguous, since $G_n(f)$ can never be negative. Now applying Schwartz's inequality, we get

$$\left|\int_{-\infty}^{\infty} S(f)H(f)e^{j2\pi fT}\,df\right|^2 \le \int_{-\infty}^{\infty} |H(f)|^2 G_n(f)df \int_{-\infty}^{\infty} \frac{|S(f)|^2}{G_n(f)}df \tag{4.92}$$

Combining this with Eq. (4.87), we have

$$\begin{aligned}\rho &\le \frac{\displaystyle\int_{-\infty}^{\infty} |H(f)|^2 G_n(f)df \int_{-\infty}^{\infty} |S(f)|^2 / G_n(f)df}{\displaystyle\int_{-\infty}^{\infty} |H(f)|^2 G_n(f)df} \\ &= \int_{-\infty}^{\infty} \frac{|S(f)|^2}{G_n(f)}df\end{aligned} \tag{4.93}$$

Equation (4.93) fixes an upper bound on the SNR at the output of the filter. If we can somehow guess at an $H(f)$ that yields this maximum, we need look no further.

Before attempting the guess, let us recap the approach we are taking. We wish to choose $H(f)$ so as to maximize Eq. (4.87). This is a difficult mathematical problem to solve. Instead of attempting a direct solution, we have placed an upper bound upon the equation. If the SNR cannot exceed that bound, and we somehow find an $H(f)$ that achieves the bound, we have solved the original problem. (In general, there is no guarantee that we can even achieve a bound of this type; however, in this case, we can.)

We have now reduced the problem to finding an $H(f)$ that reduces Eq. (4.87) to Eq. (4.93). We are asking you to be creative, and there is no road map for doing so. You need to stare at the two expressions, hoping for an inspiration. [Indeed, except by hindsight, we have no assurance that an $H(f)$ exists that will achieve the bound of Eq. (4.93).]

The answer is (if you figured this out, your insight is excellent)

$$H(f) = e^{-j2\pi fT}\frac{S^*(f)}{G_n(f)} \tag{4.94}$$

$S^*(f)$ is the complex conjugate of $S(f)$. If you were not able to see this answer, you might wish to go back and assume that the noise is white (as we shall in a moment). That is, assume that $G_n(f)$ is a constant. This makes the creative inspiration easier to achieve.

Since $H(f)$ appears as a square in both the numerator and denominator of Eq. (4.87), any scaling factor can be applied to $H(f)$ without affecting the SNR. That is, the $H(f)$ of Eq. (4.94) can be multiplied by any constant. We shall therefore rewrite this equation, inserting C for an arbitrary constant:

$$H(f) = Ce^{-j2\pi fT}\frac{S^*(f)}{G_n(f)} \tag{4.95}$$

Alas, simple amplification does not improve the SNR, since both the signal and noise are multiplied by the same amount.

Now let us assume that the input noise is white, so that $G_n(f) = N_0/2$. The matched filter of Eq. (4.95) then becomes

$$H(f) = \frac{2C}{N_0}e^{-j2\pi fT}S^*(f) \text{ "=" } Ce^{-j2\pi fT}S^*(f) \tag{4.96}$$

Note that because C is an arbitrary constant, it would create unnecessary bookkeeping to write $2C/N_0$ in Eq. (4.96). That is why we have replaced $2C/N_0$ with C. (We put the equals sign in quotes so that you don't draw the conclusion that C must be zero.)

We can find the SNR at the output of the matched filter (with white noise at the input) directly from Eq. (4.93):

$$\rho = \frac{2}{N_0}\int_{-\infty}^{\infty}|S(f)|^2\,df = \frac{2}{N_0}\int_{-\infty}^{\infty}s^2(t)\,dt \tag{4.97}$$

The last equality in Eq. (4.97) results from Parseval's theorem.

The inverse Fourier transform of Eq. (4.96) yields the impulse response of the matched filter:

$$h(t) = Cs(T - t) \tag{4.98}$$

This is found by noting that the inverse transform of $S^*(f)$ is $s(-t)$ and the exponential leads to a time shift. At this point, there is no assurance that this filter is physically realizable (i.e., causal).

Example 4.19

Find the impulse response of the matched filter for the two functions of time shown in Fig. 4.30.

Solution: $h(t)$ is derived directly from Eq. (4.98). The result is shown in Fig. 4.31.

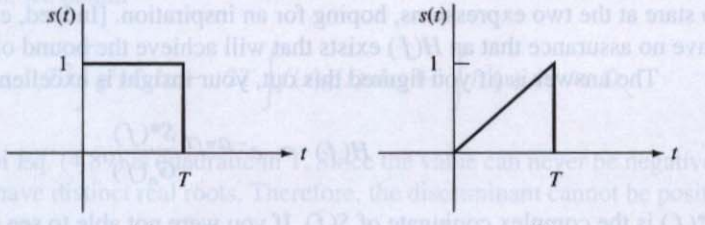


Figure 4.30 Functions of time for Example 4.19.

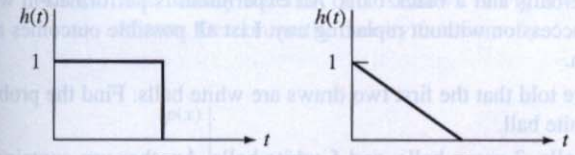


Figure 4.31 Matched filter for Example 4.19.

The actual function of time at the output of the matched filter can be found by convolving the input function with the impulse response. Therefore,

$$s_o(t) + n_o(t) = [s(t) + n(t)] * h(t)$$

and at time $t = T$, we have

$$s_o(T) + n_o(T) = \int_0^T [s(\tau) + n(\tau)]s(\tau)\, d\tau$$

Thus, the matched filter is equivalent to the system of Fig. 4.32.

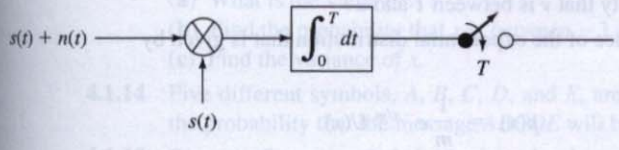


Figure 4.32 Correlator.

The operation being performed by this system is called *correlation* (i.e., multiply two functions of time together and integrate the product). For that reason, the matched filter is often referred to as a *correlator*. In a generalized sense, the filter is finding the projection of the input signal in the direction of $s(t)$. Since the system is aligned in that direction, the output SNR is maximized.

Example 4.20

Find the output SNR of a matched filter, where the signal is

$$s(t) = A, \quad \text{for} \quad 0 < t < T$$

The noise is white with power spectral density $N_0/2$.

Solution: The matched filter achieves the SNR of Eq. (4.97). Therefore,

$$\text{SNR} = \frac{2}{N_0}\int_{-\infty}^{\infty} s^2(t)\, dt = \frac{2A^2T}{N_0}$$

PROBLEMS

4.1.1 Three coins are tossed at the same time. List all possible outcomes of this experiment. List five representative events. Find the probabilities of the following events: {all tails}; {one head only}; {three matches}

4.1.2 Find the probability that, if a perfectly balanced die is rolled, the number of spots on the face turned up is greater than or equal to 2.

4.1.3 An urn contains 4 white balls and 7 black balls. An experiment is performed in which 3 balls are drawn out in succession without replacing any. List all possible outcomes and assign probabilities to each.

4.1.4 In Problem 4.1.3, you are told that the first two draws are white balls. Find the probability that the third is also a white ball.

4.1.5 An urn contains 4 red balls, 7 green balls, and 5 white balls. Another urn contains 5 red balls, 9 green balls, and 2 white balls. One ball is drawn from each urn. What is the probability that both balls will be of the same color?

4.1.6 Three people, A, B, and C, live in the same neighborhood and use the same bus line to go to work. Each of the three has a probability of $\frac{1}{4}$ of making the 6:10 bus, a probability of $\frac{1}{2}$ of making the 6:15 bus, and a probability of $\frac{1}{4}$ of making the 6:20 bus. Assuming independence, what is the probability that they all take the same bus?

4.1.7 The probability density function of a certain voltage is given by

$$p_v(v) = ve^{-v}U(v)$$

where $U(v)$ is the unit step function.

(a) Sketch this probability density function.

(b) Sketch the distribution function of v.

(c) What is the probability that v is between 1 and 2?

4.1.8 Find the mean and variance of the exponential distribution that is given by

$$p(x) = \frac{1}{m}e^{-x/m}U(x)$$

4.1.9 Find the density of $y = |x|$, given that $p_x(x)$ is as shown in Fig. P4.1.9. Also, find the mean and variance of both y and x.

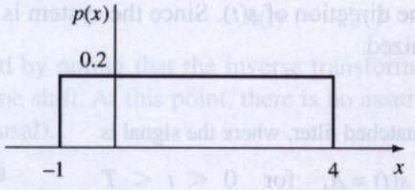


Figure P4.1.9

4.1.10 Find the mean and variance of x, where the density of x is as shown in Fig. P4.1.10.

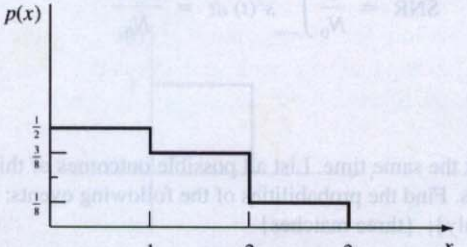


Figure P4.1.10

4.1.11 The density function of x is shown in Fig. P4.1.11. A random variable y is related to x as shown. Determine the density function of y.

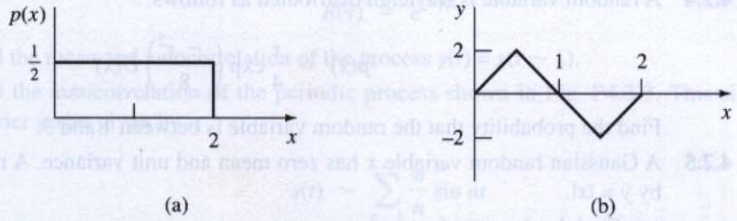


Figure P4.1.11

4.1.12 Find the expected value of $y = \sin x$ if x is uniformly distributed between 0 and 2π.

4.1.13 A random variable x has the probability density function

$$p_x(x) = Ae^{-2|x|}$$

(a) What is the value of A?
(b) Find the probability that x is between -3 and $+3$.
(c) Find the variance of x.

4.1.14 Five different symbols, A, B, C, D, and E, are transmitted with equal probability. What is the probability that the message $ABCDE$ will be received?

4.1.15 Binary information is being received, where the probabilities of 0 and 1 are equal. Received messages are divided into 5-bit words. What is the probability that the first received message will contain at least one zero?

4.1.16 A random variable x has the probability density shown in Fig. P4.1.16. A new variable y is defined as the magnitude of x:

$$y = |x|$$

(a) Find the probability density of y, $p(y)$.
(b) Find the mean value of y.
(c) Find the variance of y.

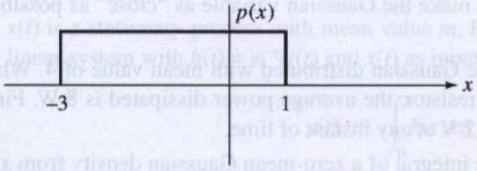


Figure P4.1.16

4.2.1 A Gaussian random variable has a mean value of 2. The probability that the variable lies between 2 and 5 is 0.3. Find the probability that the variable is between 1 and 3.

4.2.2 A Gaussian random variable has a variance of 9. The probability that the variable is greater than 5 is 0.1. Find the mean of the random variable.

4.2.3 A zero-mean Gaussian random variable has a variance of 4. Find x_0 such that

$$\Pr\{|x| > x_0\} < 0.001$$

4.2.4 A random variable is Rayleigh distributed as follows:

$$p(x) = \frac{x}{4}\exp\left(\frac{-x^2}{8}\right)U(x)$$

Find the probability that the random variable is between 1 and 3.

4.2.5 A Gaussian random variable x has zero mean and unit variance. A new variable is defined by $y = |x|$.

(a) Find the density of y.
(b) Find the mean value of y.
(c) Find the variance of y.

4.2.6 A function is defined by

$$y = e^{-2x}$$

Find the density of y if

(a) x is uniformly distributed between 0 and 3.
(b) $p(x) = e^{-x}U(x)$

4.2.7 You are told that the mean rainfall per year in California is 10 inches and that the standard deviation in the amount of rain per year is 1 inch. Can you tell, from this information, what is the density of rainfall in California? If so, roughly sketch this density function.

4.2.8 The random variable x is Rayleigh distributed. Find the mean, second moment, and variance.

4.2.9 The random variable x is uniformly distributed between -1 and $+1$. The random variable y is uniformly distributed between 0 and $+5$. x and y are independent of each other. A new variable, z, is formed as the sum of x and y.

(a) Find the probability density function of z.
(b) Find the mean and variance of x, y, and z.
(c) Find a general relationship among the means of x, y, and z.
(e) Repeat part (c) for the variance.

4.2.10 **(a)** Find and sketch the density of the sum of two independent random variables as follows: One of the variables is uniformly distributed between -1 and $+1$; the second variable is triangularly distributed between -2 and $+2$.
(b) Compare the result of part (a) to a Gaussian density function, where the variance should be chosen to make the Gaussian variable as "close" as possible to your answer to part (a).

4.2.11 A voltage is known to be Gaussian distributed with mean value of 4. When this voltage is impressed across a 4-Ω resistor, the average power dissipated is 8 W. Find the probability that the voltage exceeds 2 V at any instant of time.

4.2.12 **(a)** You are told that the integral of a zero-mean Gaussian density from $x = 10$ to $x = \infty$ is equal to 0.02. Find the variance of this random variable.
(b) Find the probability that x is between 1 and 3.

4.2.13 A Gaussian random variable x has a mean value of m and a variance of 4. You are told that the probability that the variable is greater than 6 is 0.01. Find the mean value, m.

4.3.1 A random variable k is uniformly distributed in the interval between -1 and $+1$. Sketch several possible samples of the process

$$x(t) = k\sin 2\pi t$$

4.3.2 $x(t)$ is a stationary process with mean value 1 and autocorrelation

$$R(\tau) = e^{-|\tau|}$$

Find the mean and autocorrelation of the process $y(t) = x(t - 1)$.

4.3.3 Find the autocorrelation of the periodic process shown in Fig. P4.3.3. This signal has a Fourier series given by

$$s(t) = \sum_{n=1}^{\infty} \frac{\pi}{n} \sin nt$$

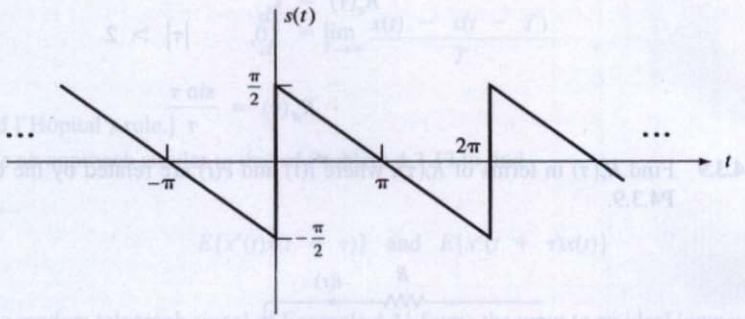


Figure P4.3.3

4.3.4 $x(t)$ is a stationary process with zero mean value. Is the process

$$y(t) = tx(t)$$

stationary? Find the mean and autocorrelation of $y(t)$.

4.3.5 Given a stationary process $x(t)$, find the autocorrelation of

$$y(t) = x(t - 1) + \sin 2\pi t$$

in terms of $R_x(\tau)$.

4.3.6 $x(t)$ is a stationary process with mean value m. Find the mean value of the output $y(t)$ of a linear system with $h(t) = e^{-t}U(t)$ and $x(t)$ as input. That is,

$$y(t) = \int_0^t h(\tau)x(t - \tau)\, d\tau$$

4.3.7 A random process is defined by

$$x(t) = K_1 t + K_2$$

where K_2 is a deterministic constant and K_1 is uniformly distributed between 0 and 1.

(a) Sketch several samples of this process.

(b) Find the mean of the process.

(c) Write an expression for the autocorrelation of the process.
(d) Is the process stationary?
(e) If K_1 is now uniformly distributed between -1 and $+1$, what changes occur in your answers to parts (b), (c), and (d)?

4.3.8 Which of the following could *not* be the autocorrelation function of a process?

$$R_a(\tau) = \begin{cases} 1 - |\tau|, & |\tau| < 1 \\ 0, & |\tau| > 1 \end{cases}$$

$$R_b(\tau) = 5\sin 3\tau$$

$$R_c(\tau) = \begin{cases} 1 & |\tau| < 2 \\ 0, & |\tau| > 2 \end{cases}$$

$$R_d(\tau) = \frac{\sin \tau}{\tau}$$

4.3.9 Find $R_e(\tau)$ in terms of $R_i(\tau)$, where $i(t)$ and $e(t)$ are related by the circuit shown in Fig. P4.3.9.

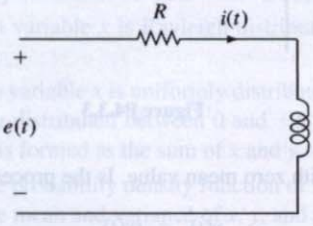


Figure P4.3.9

4.3.10 Given a constant a and a random variable f with density $p_f(f)$, form

$$x(t) = ae^{j2\pi ft}$$

Find $R_x(\tau)$, and show that

$$G_x(f) = |a|^2 p_f(f)$$

4.3.11 Find the autocorrelation and power spectral density of the wave

$$v(t) = A\cos(2\pi f_c t + \theta)$$

where A and f_c are not random and θ is uniformly distributed between 0 and 2π. [*Hint:* Use the definitions of autocorrelation and expected value.]

4.3.12 You are given the *RC* circuit of Fig. P4.3.12 with input function as shown. You wish to choose the value of R so that the total energy at the output is 50% of the input energy.

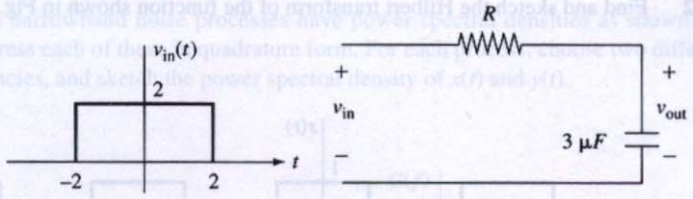


Figure P4.3.12

4.3.13 Use the result of Example 4.8 to find the autocorrelation of the process $y(t) = dx/dt$ in terms of the autocorrelation of $x(t)$. [*Hint:* You can use the definition of the derivative,

$$\frac{dx}{dt} = \lim_{T \to \infty} \frac{x(t) - x(t - T)}{T}$$

and l'Hôpital's rule.]

4.3.14 Use an approach similar to that of Problem 4.3.13 to find

$$E\{x'(t)x(t + \tau)\} \quad \text{and} \quad E\{x'(t + \tau)x(t)\}$$

4.3.15 The random telegraph signal of Example 4.11 forms the input to an ideal lowpass filter with cutoff at f_m. Find the ratio of output to input power as function of f_m and λ.

4.4.1 $x(t)$ is white noise with autocorrelation $R_x(\tau) = \delta(\tau)$. It forms the input to an ideal lowpass filter with cutoff frequency f_m. Referring to Fig. P4.4.1, find the average power of the output signal.

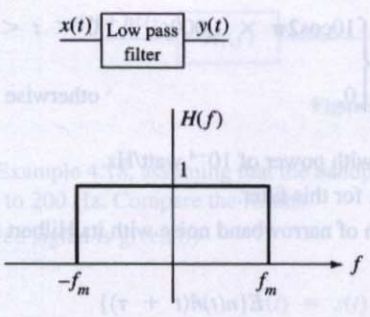


Figure P4.4.1

4.5.1 Prove that the inverse Fourier transform of a Hilbert transform is

$$\mathcal{F}^{-1}(-j\text{sgn}(f)) = \frac{1}{\pi t}$$

4.5.2 Find and sketch the Hilbert transform of the function shown in Fig. P4.5.2.

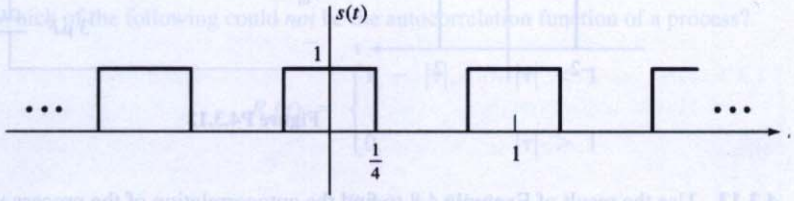


Figure P4.5.2

4.5.3 Express the narrowband noise processes shown in Fig. P4.5.3 in quadrature form. For each process, choose two different center frequencies, and sketch the power spectral density of $x(t)$ and $y(t)$.

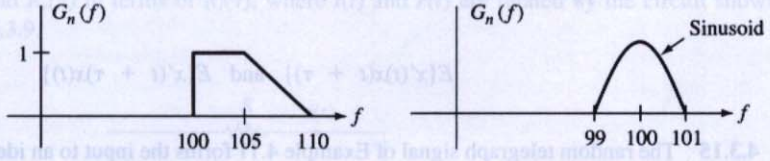


Figure P4.5.3

4.5.4 **(a)** Find the matched filter for the signal

$$s(t) = \begin{cases} 10\cos 2\pi \times 1{,}000t, & 0 < t < 50\ \mu\text{sec} \\ 0, & \text{otherwise} \end{cases}$$

in white Gaussian noise with power of 10^{-4} watt/Hz.

(b) Find the output SNR for this filter.

4.5.5 Find the cross correlation of narrowband noise with its Hilbert transform. That is, evaluate

$$E\{n(t)\hat{n}(t + \tau)\}$$

in terms of the autocorrelation of $n(t)$.

4.5.6 Show that the cross correlation between the in-phase and quadrature terms in a narrowband noise expansion is given by

$$R_{xy}(\tau) = R_n(\tau)\sin 2\pi f_c\tau - \hat{R}_n(\tau)\cos 2\pi f_c\tau$$

4.5.7 Two narrowband noise processes have power spectral densities as shown in Fig. P4.5.7. Express each of these in quadrature form. For each process, choose two different center frequencies, and sketch the power spectral density of $x(t)$ and $y(t)$.

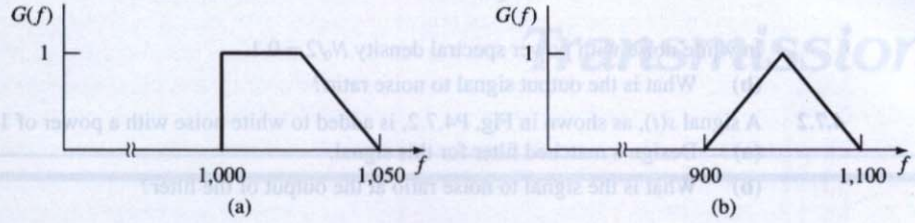


Figure P4.5.7

4.6.1 In a given communication system, the signal is $s(t) = 20\cos 2\pi t$. Noise of power spectral density $G_n(f) = e^{-3|f|}$ is added to the signal, and the resultant sum forms the input to a filter.
(**a**) Find the SNR at the input to the filter.
(**b**) If the filter is ideal lowpass with a cutoff of 2 Hz, find the improvement in the SNR, and express it in dB.
(**c**) If the filter is ideal bandpass with a passband from 0.9 to 1.1 Hz, find the improvement in SNR, and express it in dB.

4.6.2 You are given the system shown in Fig. P4.6.2, which is composed of two cascaded filters. $s(t)$ and $n(t)$ are as given in Problem 4.6.1. $H_1(f)$ is a lowpass filter with cutoff of 2 Hz. $H_2(f)$ is a bandpass filter with passband from 0.9 to 1.1 Hz. Find the improvement in SNR of $H_1(f)$ and of $H_2(f)$, and express these in dB. Now find the overall improvement in SNR of the cascaded system, and compare it to the individual improvement in SNR of the two filters.

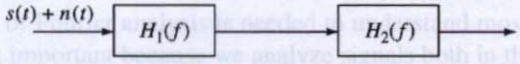


Figure P4.6.2

4.6.3 Repeat Example 4.18, assuming that the bandpass filter passes from 90 to 210 Hz instead of from 10 to 200 Hz. Compare the results.

4.6.4 A received signal is given by

$$r(t) = s(t) + n(t)$$

where

$$s(t) = 5\cos 2\pi \times 3{,}000t + 15\cos 2\pi \times 2{,}100t$$

The noise $n(t)$ is white with power $N_0 = 0.05$ watt/Hz. The signal $r(t)$ is put through a bandpass filter with a passband between 1,900 and 2,200 Hz. Find the signal to noise ratio at the output of the filter.

4.7.1 (a) Find the matched filter for the signal

$$s(t) = \begin{cases} 5\cos 2\pi \times 2{,}000t, & 0 < 1 < 0.005 \\ 0, & \text{otherwise} \end{cases}$$

in white noise with power spectral density $N_0/2 = 0.1$.

(b) What is the output signal to noise ratio?

4.7.2 A signal $s(t)$, as shown in Fig. P4.7.2, is added to white noise with a power of 10^{-2} watt/Hz.

(a) Design a matched filter for this signal.

(b) What is the signal to noise ratio at the output of the filter?

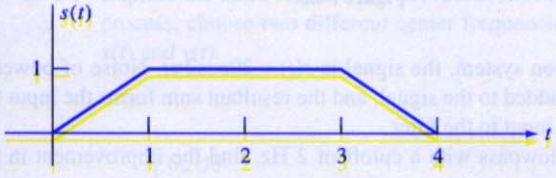


Figure P4.7.2

5 Baseband Transmission

5.0 PREVIEW

What We Will Cover and Why You Should Care

You have finally arrived at the first chapter dealing with communication. The previous chapters have simply laid the necessary groundwork.

In this chapter, we deal with what is known as *baseband transmission*. The term *baseband* refers to the frequency content of the signal. The frequencies used to transmit baseband signals are relatively low. In Chapters 6 and 7, we explore techniques of sending signals in which the transmitted frequencies are high.

Baseband transmission is used in a variety of communication systems, from telephone loops to intercoms. Even if the type of communication used is not baseband, it is important to understand baseband transmission as a stepping-stone to other techniques.

Necessary Background

A working knowledge of Fourier analysis is needed to understand most analog communication systems. This is important because we analyze signals both in the time domain and in the frequency domain.

Analysis of the performance of systems (Section 5.5) requires random process analysis.

5.1 ANALOG BASEBAND

When we use the term *analog baseband*, we are referring to analog signals with Fourier transforms occupying frequencies extending to zero (dc). The reason for using the term *baseband* is that the frequencies are *not* shifted to some other nonzero point on the frequency axis. In later chapters, we explore techniques for shifting frequencies to a range centered around a nonzero frequency.

Since baseband signals occupy relatively low frequencies, they are not suited for transmission through bandpass channels. Baseband signals are typically transmitted through wires or cables.

Because telephones form the backbone of traditional communication systems, the transmission of voice signals represents the most prevalent application of baseband analog communication. We therefore concentrate initially on voice transmission.

The human ear is capable of hearing signals with frequencies in the range of about 20 Hz to 20 kHz. In fact, most people cannot hear the upper portion of this range, and a particular person's upper frequency cutoff decreases with age. It is probably no accident of evolution that signals generated by human beings and their vocal cords fall within the audible range. The magnitude of the Fourier transform of a typical speech waveform is shown in Fig. 5.1. The location of the peak of the waveform depends on the physiology of the speaker (i.e., the resonant frequency of the vocal cavity). It also depends on what the person is saying and what language is being spoken. If, for example, the speaker whistles, the speech waveform is a pure sinusoid and has a Fourier transform consisting of an impulse at the frequency of the whistling. If the person hums at the same frequency, the Fourier transform contains the fundamental frequency plus harmonics at multiples of that frequency.

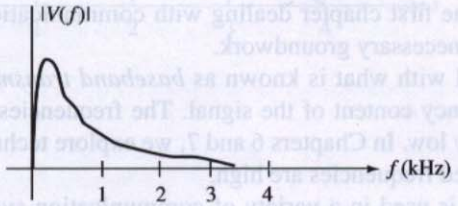


Figure 5.1 Fourier transform of typical voice waveform.

In the early days of telephony, experimentation showed that the portion of the speech waveform between about 300 Hz and 3.3 kHz was sufficient both for intelligibility of speech and for recognition of the speaker. That is, although this range is not considered high fidelity, it permits both understanding of what is being said and identification of the person saying it. High-quality music requires the presence of frequencies higher than 3.3 kHz—perhaps as high as 15 kHz or more. (AM radio transmits frequencies up to 5 kHz, while FM radio transmits frequencies up to 15 kHz. The typical high-quality home entertainment system responds to frequencies above 20 kHz.)

If you speak into a microphone and amplify the resulting waveform using electronic circuitry, you have effectively built a simple analog baseband transmitter. If the output of the transmitter is now connected to a wire channel, and the other end of the channel is connected to a loudspeaker (perhaps through some amplification if the channel contains loss), a complete baseband communication system results. Such systems are used in wire intercoms.

5.2 THE SAMPLING THEOREM

A *discrete* signal is a signal that is *not* continuous in time. That is, it has values only at disconnected points of the time axis. If the discrete signal is analog, its values at any time it is defined lie within a continuum of possible values.

We wish to find a way of converting a continuous (not discrete) analog waveform into a discrete signal. To do this, the time axis must somehow be made discrete. The conversion of the continuous time axis into a discrete axis is accomplished by time sampling. The *sampling theorem* states the following:

> If the Fourier transform of a function of time is zero for $f > f_m$ and the values of the function are known for $t = nT_s$ (for all integer values of n), then the function is known exactly for *all* values of t.

This is remarkable: Knowing the value of the function of time at discrete time points allows us to fill in the curve between these points *precisely and accurately!* Of course, something this remarkable must have some limitations: You couldn't be given two values separated by hours and expect to fill in the curve between these points. Indeed, for the samples to give all of the information, they must be "close enough" to each other. The restriction is that the spacing T_s between samples be less than $1/2f_m$, where f_m is the maximum frequency of the signal. Alternatively, $s(t)$ can be uniquely determined from its values at a sequence of equidistant points in time. The upper limit of T_s, $1/2f_m$, is known as the *Nyquist sampling interval.*

The upper limit on T_s can be expressed in a more meaningful way by taking the reciprocal of T_s to obtain the sampling frequency, denoted $f_s = 1/T_s$ in samples per second. The restriction then becomes

$$T_s < \frac{1}{2f_m}$$

$$\frac{1}{T_s} > 2f_m$$

$$f_s > 2f_m$$

Thus, the sampling frequency must be greater than twice the highest frequency of the signal being sampled. For example, if a voice signal has 4 kHz as a maximum frequency, it must be sampled at least 8,000 times per second to comply with the conditions of the sampling theorem. Twice the highest frequency is known as the *Nyquist frequency.*

Before going further, let us observe that the spacing between the sample points is inversely related to the maximum frequency f_m. This is intuitively satisfying, since the higher f_m, the faster we would expect the function to vary. The faster the function varies, the closer together the sample points should be in order to permit reconstruction of the function.

We present two proofs of the sampling theorem. The first is physical and intuitive, while the second is more mathematical.

Proof 1. Figure 5.2 shows a pulse train multiplying the original signal $s(t)$. If the pulse train consists of narrow pulses, one would say that the output of the multiplier is a *sampled version* of the original waveform. In actuality, the output depends not only on the sample values of the input, but on a range of values around each sample point. The theory does not require these extra values, which represent redundant information. However, practical systems sometimes sample over a small range of time surrounding the actual sample points. As we prove the theorem, it should become obvious that the multiplying

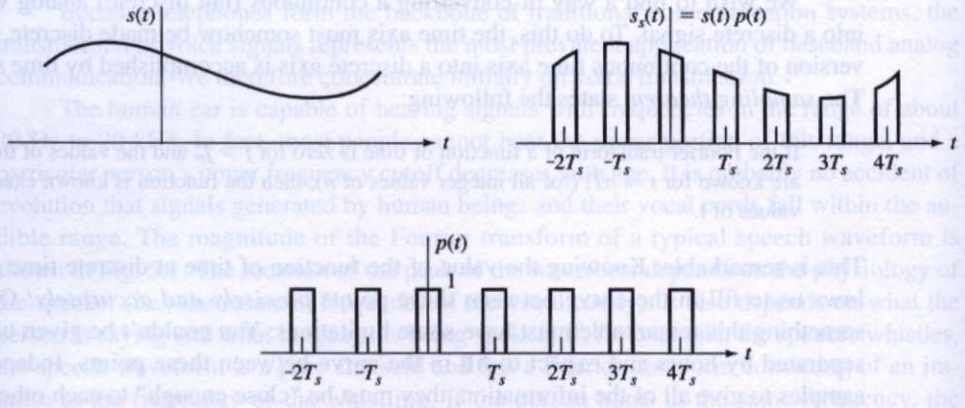


Figure 5.2 Product of pulse train and $s(t)$.

function need not consist of perfect square pulses. In fact, the function can be any periodic signal, and the pulse widths can approach zero (i.e., multiply by a train of impulses).

Multiplying $s(t)$ by $p(t)$ of the type shown in Fig. 5.2 is a form of *time gating*. It can be viewed as the opening and closing of a gate, or switch.

Our goal is to show that the original signal can be recovered from the sampled waveform $s_s(t)$. We do this by examining the Fourier transform of $s_s(t)$. The sampling theorem requires that we assume that $s(t)$ has no frequency components above f_m. The Fourier transform of $s(t)$, $S(f)$, therefore cuts off at f_m. Figure 5.3 shows a representative shape for this transform. While we use this triangular shape throughout the text, we do not mean to restrict the actual transform to that shape.

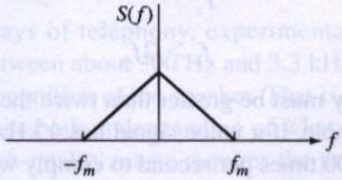


Figure 5.3 Representative $S(f)$.

Since the multiplying pulse train is assumed to be periodic, it can be expanded in a Fourier series. The $p(t)$ shown is an even function, so we can use a trigonometric series containing only cosine terms (although this is not necessary to prove the theorem). Thus,

$$\begin{aligned} s_s(t) &= s(t)p(t) \\ &= s(t)\left[a_0 + \sum_{n=1}^{\infty} a_n \cos 2\pi n f_s t\right] \\ &= a_0 s(t) + \sum_{n=1}^{\infty} a_n s(t) \cos 2\pi n f_s t \end{aligned} \tag{5.1}$$

where

$$f_s = \frac{1}{T_s} \tag{5.2}$$

The goal is to isolate the first term in the final expression of Eq. (5.1), which is proportional to the original $s(t)$. We can undo the effects of any constant multiplier with an amplifier or attenuator.

Each of the terms in the summation of Eq. (5.1) is of the form of $s(t)$ multiplied by a sinusoid. When a time signal is multiplied by a sinusoid, the result is a shift of all frequencies of the signal by an amount equal to the frequency of the sinusoid. The frequency content of each term in Eq. (5.1) is then centered around the frequency of the multiplying sinusoid. (When we discuss AM, we will call this the *carrier frequency*.) The Fourier transform of $s_s(t)$ is sketched in Fig. 5.4.

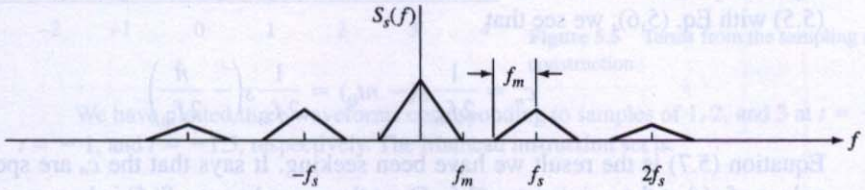


Figure 5.4 Fourier transform of sampled wave.

The shape centered at the origin is the transform of $a_0 s(t)$, and the shifted versions represent the transforms of the various harmonic terms. We see that the various terms do not overlap in frequency, provided that $f_s > 2f_m$. But this is nothing more than the condition given in the sampling theorem. Since the various terms occupy different bands of frequency, they can be separated from each other using linear filters. A lowpass filter with a cutoff frequency of f_m can be used to recover the $a_0 s(t)$ term.

Proof 2. The second proof we present is less intuitive than the first. We take the time to explore it, since the approach supplies an insight into the mathematical principles of sampling.

Since $S(f)$ is nonzero along a finite portion of the f-axis, we can expand it in a Fourier series in the interval

$$-f_m < f < f_m$$

In expanding $S(f)$ in this manner, you should be careful not to let the change in notation confuse the issue. The t used in the Fourier series is an independent functional variable, and any other letter could be substituted for it. Performing the Fourier series expansion, we obtain

$$S(f) = \sum_{n=-\infty}^{\infty} c_n e^{jn2\pi t_0 f} \tag{5.3}$$

where

$$t_0 = \frac{1}{2f_m} \tag{5.4}$$

The c_n in Eq. (5.3) are given by

$$c_n = \frac{1}{2f_m}\int_{-f_m}^{f_m} S(f)e^{-jn2\pi t_0 f}df \tag{5.5}$$

However, the Fourier inversion integral tells us that

$$s(t) = \int_{-\infty}^{\infty} S(f)e^{j2\pi ft}df = \int_{-f_m}^{f_m} S(f)e^{j2\pi ft}df \tag{5.6}$$

In the rightmost expression in Eq. (5.6), the limits of integration have been reduced, since $S(f)$ is equal to zero outside of the interval between $-f_m$ and $+f_m$. Upon comparing Eq. (5.5) with Eq. (5.6), we see that

$$c_n = \frac{1}{2f_m}s(-nt_0) = \frac{1}{2f_m}s\left(-\frac{n}{2f_m}\right) \tag{5.7}$$

Equation (5.7) is the result we have been seeking. It says that the c_n are specified by the values of $s(t)$ at the points $t = n/2f_m$. Once the c_n are known, $S(f)$ is known, and once $S(f)$ is known, $s(t)$ is known. We have thus proven the sampling theorem.

Although the proof is complete, we will carry the mathematics a step further to actually solve for $s(t)$ in terms of the sample values. We substitute the c_n of Eq. (5.7) into Eq. (5.3) to get

$$S(f) = \frac{1}{2f_m}\sum_{n=-\infty}^{\infty} s\left(-\frac{n}{2f_m}\right)e^{jn\pi f/f_m} \tag{5.8}$$

We now find the inverse Fourier transform of $S(f)$:

$$\begin{aligned} s(t) &= \sum_{n=-\infty}^{\infty}\frac{1}{2f_m}\int_{-f_m}^{f_m} s\left(-\frac{n}{2f_m}\right)e^{jn\pi f/f_m}e^{j2\pi ft}df \\ &= \sum_{n=-\infty}^{\infty} s\left(-\frac{n}{2f_m}\right)\left[\frac{\sin(2\pi f_m t + n\pi)}{2\pi f_m t + n\pi}\right] \end{aligned} \tag{5.9}$$

Equation (5.9) is the final statement of the sampling theorem. We can use it to find the value of $s(t)$ at any point in time simply by knowing the sample values of $s(t)$. That is, the only unknowns on the right side of the equation are the sample values.

You can get a feel for the sampling theorem by sketching a few terms in Eq. (5.9). We do this in Figure 5.5 for a representative $s(t)$ and three sample points. Note that only the term centered at each sampling point has non-zero value; all of the other components go to zero at the sampling points. Between sampling points, we must calculate the sum of the various terms from adjacent points.

Computer Exercise

The waveform shown in Fig. 5.5 was generated using computer software. We present the instruction set for both *Mathcad* and MATLAB.

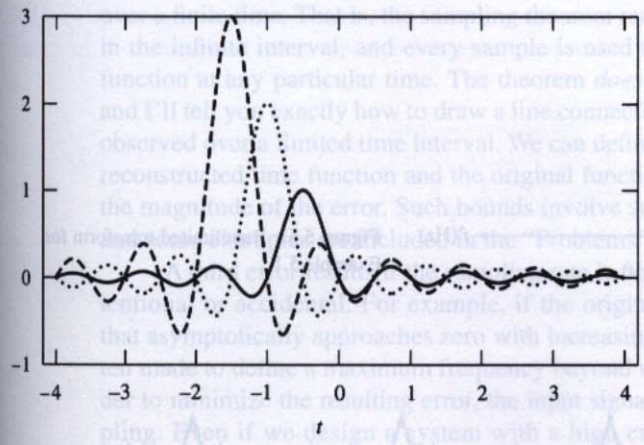


Figure 5.5 Terms from the sampling reconstruction.

We have plotted three waveforms corresponding to samples of 1, 2, and 3 at $t = -0.5$, $t = -1$, and $t = -1.5$, respectively. The Mathcad instruction set is:

```
t:=-4,-3.99..4
s1(t):=sin(2*π*t)+π/2*π*t+π
s2(t):=2*sin(2*π*t+2*π)/2*π*t+2*π
s3(t):=3*sin(2*π*t+3*π)/2*π*t+3*π
```

Notes: Enter ":=" by simply pressing ":". Enter ".." by pressing ";". Enter π by pressing CONTROL+p.

Then press "@" to create a graph. Enter "t" for the abscissa and "s1(t), s2(t), s3(t)" for the ordinate.

The MATLAB instruction set is:

```
t= -4:.01:4;
s1=sin(2*pi*t+pi)./(2*pi*t+pi);
s2=2*sin(2*pi*t+2*pi)./(2*pi*t+2*pi);
s3=3*sin(2*pi*t+3*pi)./(2*pi*t+3*pi);
plot(t,s1,t,s2,t,s3)
```

Notes: The semicolon (;) after each instruction lines stops MATLAB from printing all results immediately after pressing RETURN.

The period in front of the division sign is critical. If it is omitted, MATLAB presupposes that it is dividing two matrices (vectors).

The plot statement superimposes three graphs on the same set of axes.

Example 5.1

A bandlimited signal occupies the frequency range between 990 Hz and 1,010 Hz. A typical Fourier transform is shown in Figure 5.6. Although the sampling theorem indicates that the sampling rate must be higher than 2,020 samples per second, investigate the possibilities of sampling at rates as low as 20 samples per second.

Solution: Figure 5.7 shows the Fourier transform that results from sampling at twice the highest frequency (2,020 samples per second) and also at twice the bandwidth (20 samples per second). We are assuming that all harmonics have equal amplitude (i.e., we assume sampling with an ideal impulse train).

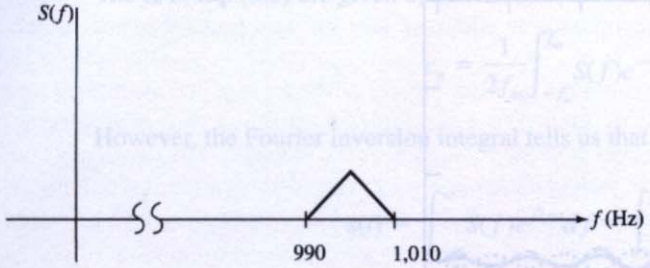
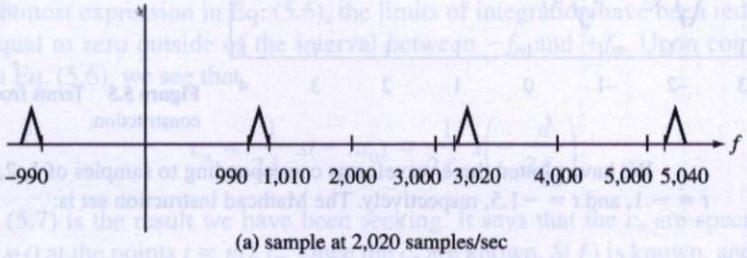
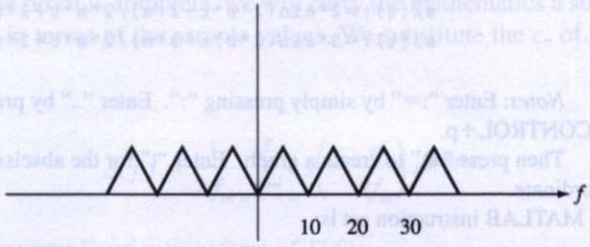


Figure 5.6 Bandlimited transform for Example 5.1.



(a) sample at 2,020 samples/sec



(b) sample at 20 samples/sec

Figure 5.7 Fourier transform of sampled waveforms.

Note that it is possible to recover the original signal when sampling at only 20 samples per second, *provided that* the exact band of frequencies occupied is known. At this lower sampling rate, it is not sufficient to know only the maximum frequency. Note also that if the limits of the band were not multiples of the bandwidth, we would need to sample at a higher rate to avoid aliasing. (We explore this phenomenon in the "Problems" section at the end of this chapter.)

The preceding example shows that it is possible to recover a bandlimited signal by sampling at a rate as low as twice the *bandwidth* of the signal.

Errors in Sampling The sampling theorem indicates that $s(t)$ can be perfectly recovered from its samples. If sampling is attempted in the real world, however, errors result from three major sources. *Round-off errors* occur when the various sample values are rounded off in the communication system. Rounding off takes place in digital communication, where we send only discrete values. We will later call this error *quantization noise*, and it is examined in detail in Chapter 8. *Truncation errors* occur if the sampling is done

over a finite time. That is, the sampling theorem requires that samples be taken for all time in the infinite interval, and every sample is used to reconstruct the value of the original function at any particular time. The theorem *does not* say, "Give me two sample values, and I'll tell you exactly how to draw a line connecting them." In a real system, the signal is observed over a limited time interval. We can define an error as the difference between the reconstructed time function and the original function, and upper bounds can be placed on the magnitude of the error. Such bounds involve sums of the rejected time sample values, and some examples are included in the "Problems" section at the end of the chapter.

A third error results if the sampling rate is not high enough. This situation can be intentional or accidental. For example, if the original time signal has a Fourier transform that asymptotically approaches zero with increasing frequency, a conscious decision is often made to define a maximum frequency beyond which signal energy is negligible. In order to minimize the resulting error, the input signal is often lowpass filtered prior to sampling. Even if we design a system with a high enough sampling rate, an unanticipated high-frequency signal (or noise) may appear at the input. In either case, the error caused by sampling too slowly is known as *aliasing*, a name derived from the fact that the higher frequencies disguise themselves in the form of lower frequencies. This is the same phenomenon that occurs when a rotating device is viewed as a sequence of individual frames, as in a television picture. As the rotation speed of the device increases, a point is reached where the perceived angular velocity starts to decrease. Eventually, a speed is reached (matched to the frame rate) at which the device appears to be standing still. Further increases make the rotation appear to reverse direction.

Analysis of aliasing is most easily performed in the frequency domain. Before doing that, we illustrate a simple example of aliasing in the time domain. Figure 5.8 shows a sinusoid at a frequency of 3 Hz. Suppose we sample this sinusoid at four samples per second. The sampling theorem tells us that the minimum sampling rate for unique recovery is six samples per second, so four samples per second is not fast enough. The samples at the slower rate are indicated in the figure. But alas, these are the same samples that would result from a sinusoid at 1 Hz, as shown by the dashed curve. The 3-Hz signal is thus disguising itself (aliasing) as a 1-Hz signal.

The Fourier transform of the sampled wave is found by periodically shifting and repeating the Fourier transform of the original signal. If the original signal has frequency components above one-half of the sampling rate, these components *fold back* into the frequency band of interest. Thus, in Fig. 5.8, the 3-Hz signal folded back to fall at 1 Hz.

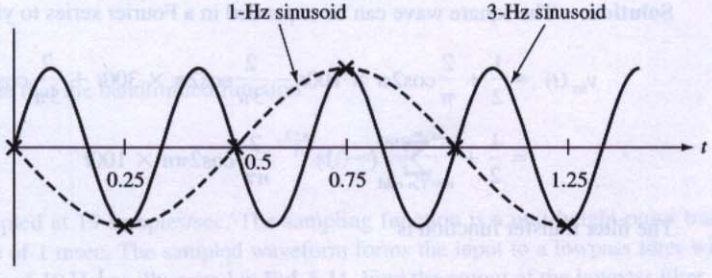


Figure 5.8 Example of aliasing.

Figure 5.9 illustrates the case where a representative signal is sampled by an ideal train of impulses (we use this as the ideal theoretical limit of narrow pulses) at less than the Nyquist rate.

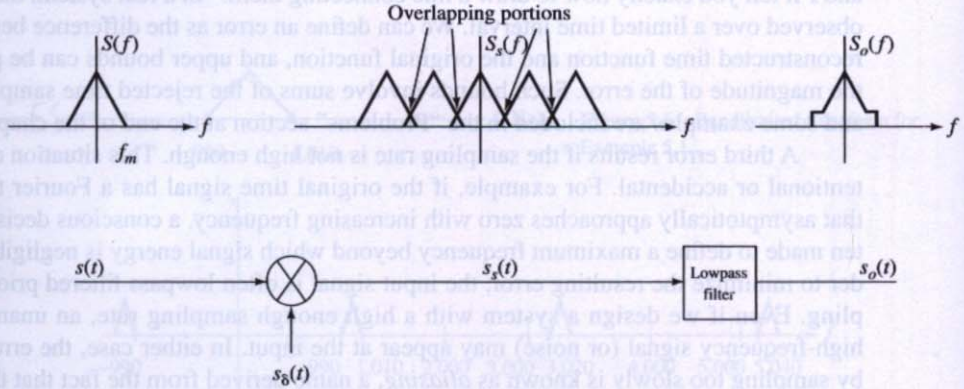


Figure 5.9 Impulse sampling at less than Nyquist rate.

Note that the transform at the output of the lowpass filter is no longer the same as the transform of the original signal. If we denote the filter output as $s_o(t)$, the error is defined as

$$e(t) = s_0(t) - s(t) \tag{5.10}$$

Taking the Fourier transform of both sides of Eq. (5.10) yields

$$\begin{aligned} E(f) &= S_0(f) - S(f) \\ &= S(f - f_s) + S(f + f_s) \text{ for } f < f_m \end{aligned} \tag{5.11}$$

Observe that if $S(f)$ were limited to frequencies below $f_s/2$, the error transform would be zero. Without assuming a specific form for $S(f)$, we cannot carry this example further. In general, various bounds can be placed upon the magnitude of the error function based upon properties of $S(f)$ for $f > f_s/2$. You can explore aliasing in more detail using computer analysis in the "problems" section at the end of the chapter.

Example 5.2

A 100-Hz square wave (assume amplitude levels of 0 and 1) forms the input to the *RC* filter shown in Figure 5.10. The output of the filter is sampled at 700 samples per second. Find the aliasing error.

Solution: The square wave can be expanded in a Fourier series to yield

$$v_{\text{in}}(t) = \frac{1}{2} + \frac{2}{\pi}\cos 2\pi \times 100t - \frac{2}{3\pi}\cos 2\pi \times 300t + \frac{2}{5\pi}\cos 2\pi \times 500t \ldots$$

$$= \frac{1}{2} + \sum_{n=1, n \text{ odd}}^{\infty} (-1)^{\frac{n+3}{2}} \frac{2}{n\pi}\cos 2\pi n \times 100t$$

The filter transfer function is

$$H(f) = \frac{1}{1 + j2\pi fRC} = \frac{1}{1 + j2\pi f(0.00167)}$$

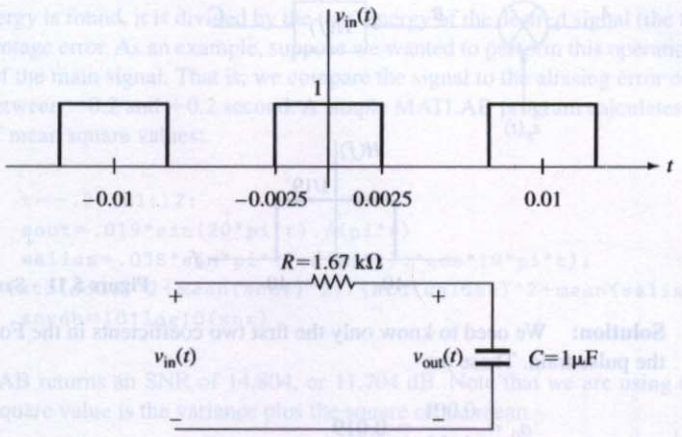


Figure 5.10 Square wave and filter for Example 5.2.

The output of the filter is found by modifying each term in the input Fourier series. The amplitude is multiplied by the magnitude of the transfer function, and the phase is shifted by the phase of the transfer function. The result is

$$v_{out}(t) = \frac{1}{2} + 0.45\cos(2\pi \times 100t - 45^\circ) - 0.067\cos(2\pi \times 300t - 71.6^\circ)$$
$$+ 0.025\cos(2\pi \times 500t - 78.7^\circ) - 0.013\cos(2\pi \times 700t - 81.9^\circ)$$

Let us assume ideal impulse sampling. The result is that the component at 500 Hz appears at 200 Hz in the reconstructed waveform, and the component at 700 Hz appears at dc (zero frequency). We shall ignore the higher harmonics. The reconstructed waveform is therefore given by

$$\frac{1}{2} + 0.45\cos(2\pi \times 100t - 45^\circ) - 0.067\cos(2\pi \times 300t - 71.6^\circ)$$
$$+ 0.025\cos(2\pi \times 200t - 78.7^\circ) - 0.013\cos(-81.9^\circ)$$

The last two terms represent the aliasing error.

Example 5.3

Assume that the bandlimited function

$$s(t) = \frac{\sin 20\pi t}{\pi t}$$

is sampled at 19 samples/sec. The sampling function is a unit-height pulse train with pulse widths of 1 msec. The sampled waveform forms the input to a lowpass filter with cutoff frequency of 10 Hz, as illustrated in Fig. 5.11. Find the output of the lowpass filter, and compare it to the original $s(t)$.

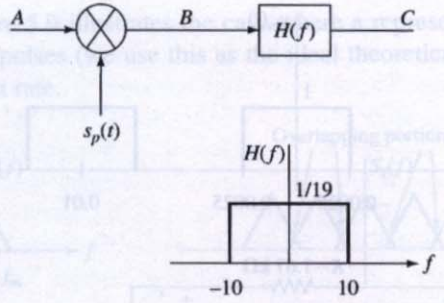


Figure 5.11 Sampling for Example 5.3.

Solution: We need to know only the first two coefficients in the Fourier series expansion of the pulse train. These are

$$a_0 = \frac{0.001}{1/19} = 0.019$$

$$a_1 = 38\int_{-5\times10^{-4}}^{5\times10^{-4}} \cos 2\pi \times 19t dt = \frac{2}{\pi}\sin(19 \times 10^{-3}\pi t) = 0.038$$

The Fourier transforms of the signals at points A, B, and C in Fig. 5.11 are shown in Fig. 5.12.

The output function of time is the inverse Fourier transform of $S_c(f)$ and is given by

$$s_0(t) = \frac{0.019\sin 20\pi t}{\pi t} + 0.038\frac{\sin\pi t}{\pi t}\cos 19\pi t$$

The second term in the result represents the aliasing error. Suppose we wish to find the maximum amplitude of this term. It should be obvious that this occurs at $t = 0$, but if you wished to use a simple MATLAB program, the instructions would be as follows:

```
t=-5:.01:=5;
s=.038*sin(pi*t)./(pi*t).*cos(19*pi*t);
MAX=max(s)
```

The maximum amplitude is 0.038 at $t = 0$, at which point the signal portion [the first term of $s_0(t)$] has amplitude 0.019. Do not be tempted to calculate a percentage error by taking the ratio of the amplitude of the error to the amplitude of the desired signal. Since the first term goes to zero at periodic points, and the second term is not necessarily zero at these same points, the percentage error would approach infinity. Errors are often analyzed by looking at the energy of the function of time representing the error. Energy is the area under the square of the function. We could therefore find the energy of the second term in the equation. Once

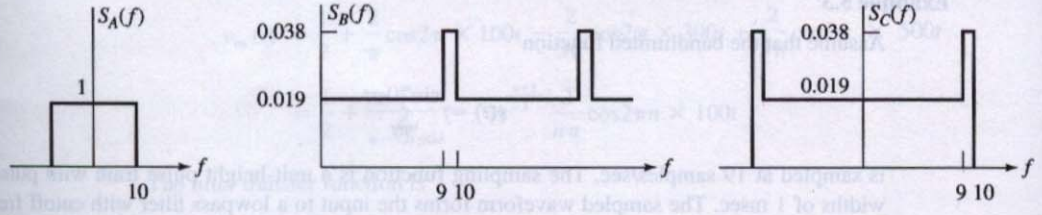


Figure 5.12 Fourier transform of sampling signals.

this energy is found, it is divided by the total energy of the desired signal (the first term) to get a percentage error. As an example, suppose we wanted to perform this operation over two side lobes of the main signal. That is, we compare the signal to the aliasing error over the range of time between -0.2 and $+0.2$ second. A simple MATLAB program calculates the SNR as the ratio of mean square values:

```
t=-.2:.01:.2;
sout=.019*sin(20*pi*t)./(pi*t)
salias=.038*sin*pi*t)./(pi*t).*cos*19*pi*t);
snr=(std(sout)^2+mean(sout)^2)/(std(salias)^2+mean(salias)^2)
snrdb=10*log10(snr)
```

MATLAB returns an SNR of 14.804, or 11.704 dB. Note that we are using the fact that the mean square value is the variance plus the square of the mean.

We conclude this section with the idea that the restriction on $S(f)$ imposed by the sampling theorem is not very severe in practice. All signals of interest in real life do possess Fourier transforms that are approximately zero above some frequency. No physical device can transmit infinitely high frequencies, since all channels contain series inductance and parallel (parasitic) capacitance. The inductance opens and the capacitance shorts as frequencies increase.

5.3 DISCRETE BASEBAND

5.3.1 Pulse Modulation

When a signal is discrete, it can be thought of as a list of numbers representing the sample values of an analog waveform. One way to send such a list through a channel is to send a pulse waveform—one pulse is placed at each sampling point. Each pulse carries information about the corresponding sample values. Each sample value can be conveyed as the amplitude, width, or position of the pulse. If we choose the amplitude of the pulse, the result is known as *pulse amplitude modulation* (PAM). Figure 5.13 illustrates a periodic pulse train $s_c(t)$, a portion of a typical analog signal $s(t)$, and the result $s_m(t)$ of controlling the pulse heights with the sample values.

Note that since the pulse tops are horizontal, the modulated waveform is *not* simply the product of the pulse train and the analog signal. Such a product would appear as in Figure 5.14. It results when $s(t)$ forms the input to a *gating* circuit.

Both of the foregoing waveforms are considered to be pulse amplitude modulated (PAM) waveforms; the waveform of Fig. 5.13 is called *flat-top* or *instantaneous-sampled* PAM, while that of Fig. 5.14 is known as *natural-sampled* PAM. The former is instantaneous sampled because the pulse height depends only upon the value of $s(t)$ at the sampling point, and not on the signal values across the range of the pulse width. Flat-top PAM is generated with a *sample-and-hold* circuit. A simplified sample-and-hold circuit is shown in Fig. 5.15.

In the figure, switch S_1 closes instantaneously at the sampling point, and the capacitor charges to the sample value. The switch is then opened, and the capacitor remains at

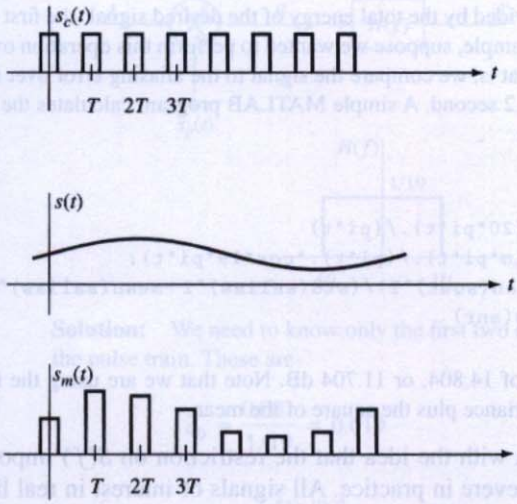


Figure 5.13 Pulse amplitude modulation.

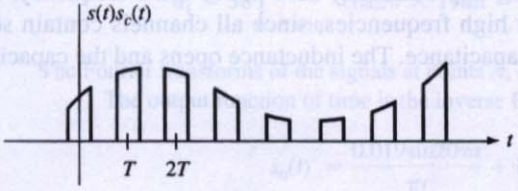


Figure 5.14 Product of pulse train and $s(t)$.

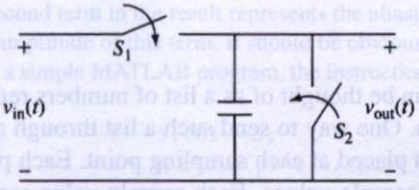


Figure 5.15 Idealized sample-and-hold circuit.

that value until the closing of switch S_2 provides a discharge path. Practical sample-and-hold circuits need additional electronics to provide the energy to charge the capacitor rapidly (i.e., the series resistance is never zero) and to prevent slow discharge (leakage) prior to switch S_2 closing.

We now calculate the Fourier transform of a PAM waveform in order to determine the channel requirements. We begin by evaluating the Fourier transform of the natural sampled waveform. The function $s_c(t)$ is expanded in a Fourier series to obtain

$$s_c(t) = a_0 + \sum_1^\infty a_n \cos 2\pi f_s t \tag{5.12}$$

When this is multiplied by $s(t)$, the result is a summation of products of $s(t)$ with sinusoids:

$$s(t)s_c(t) = a_0 s(t) + \sum_1^\infty a_n s(t) \cos 2\pi f_s t \tag{5.13}$$

The Fourier transform of each term in the summation is the signal transform $S(f)$, shifted up and down by the frequency of the sinusoid (the *modulation theorem*). The transform of $s(t)s_c(t)$ is sketched in Fig. 5.16, where we assume that $s(t)$ has the transform shown in the figure. f_m is the maximum-frequency component of $s(t)$.

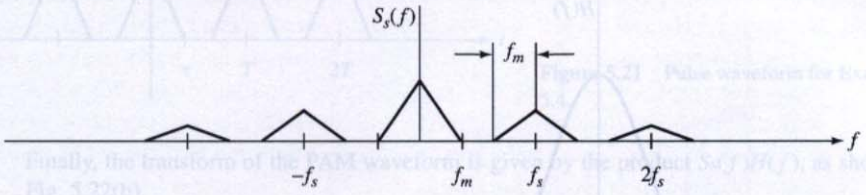


Figure 5.16 Fourier transform of natural-sampled PAM.

The Fourier transform of instantaneous-sampled PAM is more difficult to evaluate. The evaluation is simplified, however, by considering the hypothetical system of Fig. 5.17. We begin by sampling $s(t)$ with an ideal train of impulses. We then shape each impulse into the desired pulse shape—in this case, a square pulse with a flat top.

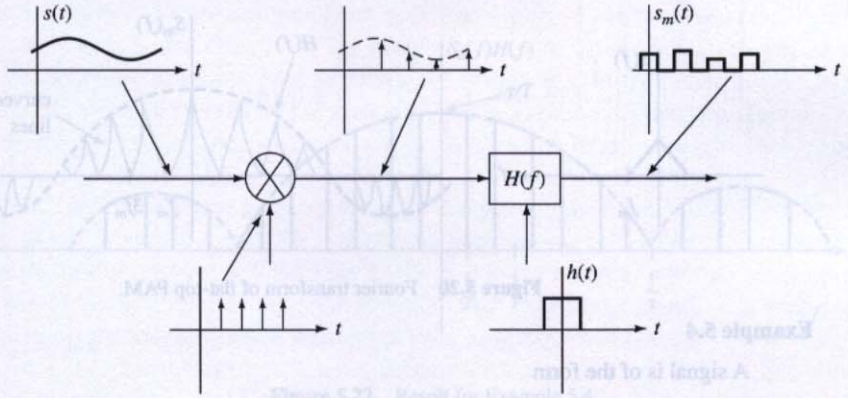


Figure 5.17 Generation of instantaneous-sampled PAM.

The Fourier transform of the sampled signal at the input to the filter is found using the sampling theorem. The Fourier series of the impulse train has equal Fourier coefficient values for all n. The Fourier transform of the impulse-sampled waveform is therefore as shown in Fig. 5.18.

The Fourier transform of the filter output (instantaneous-sampled PAM) is simply the product of the Fourier transform of the impulse-sampled waveform and the transfer

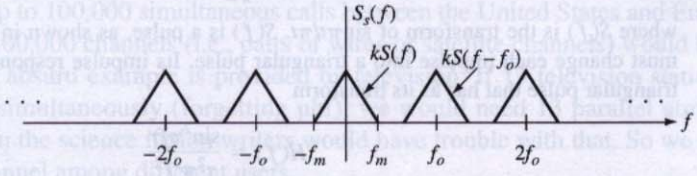


Figure 5.18 Fourier transform of impulse-sampled waveform.

function of the filter. The transfer function of the filter is shown in Fig. 5.19 and is found from a table of Fourier transforms. (See Appendix II.)

Finally, the output transform is as shown in Fig. 5.20. Note that the low-frequency portion of this transform is *not* an undistorted version of $S(f)$.

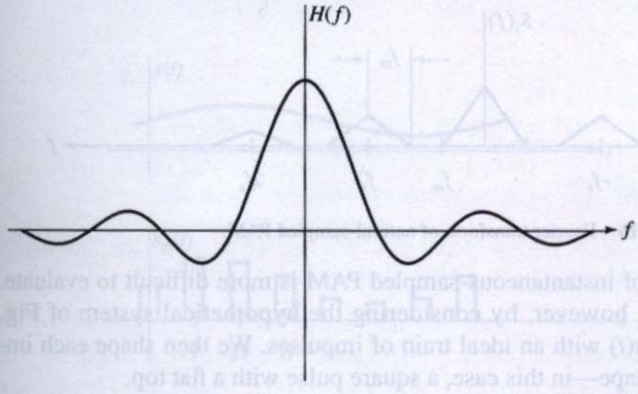


Figure 5.19 Transfer function of filter of Fig. 5.17.

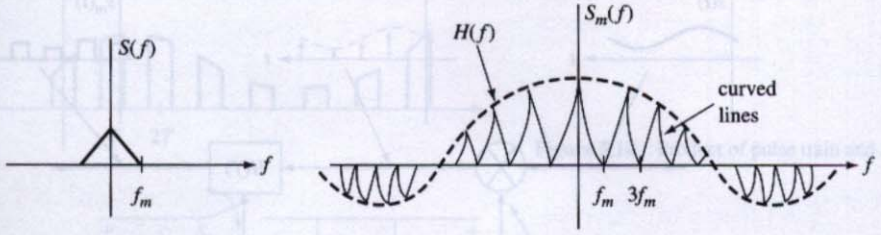


Figure 5.20 Fourier transform of flat-top PAM.

Example 5.4

A signal is of the form

$$s(t) = \frac{\sin \pi t}{\pi t}$$

It is transmitted using PAM. The pulse waveform $s_c(t)$ is a periodic train of triangular pulses, as shown in Fig. 5.21. Find the Fourier transform of the modulated waveform.

Solution: Consider the system of Fig. 5.17. The output of the ideal impulse sampler has the transform

$$S_\delta(f) = \frac{1}{T} \sum_{n=-\infty}^{\infty} S(f - nf_0)$$

where $S(f)$ is the transform of $\sin \pi t/\pi t$. $S(f)$ is a pulse, as shown in Fig. 5.22(a). The filter must change each impulse into a triangular pulse. Its impulse response is therefore a single triangular pulse that has as its transform

$$H(f) = \frac{\sin^2 \pi f \tau}{\tau^2 \pi^2 f^2}$$

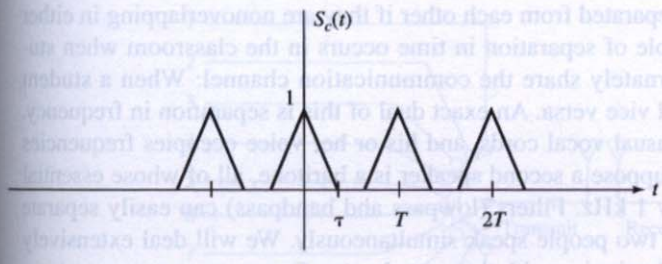


Figure 5.21 Pulse waveform for Example 5.4.

Finally, the transform of the PAM waveform is given by the product $S_\delta(f)H(f)$, as shown in Fig. 5.22(b).

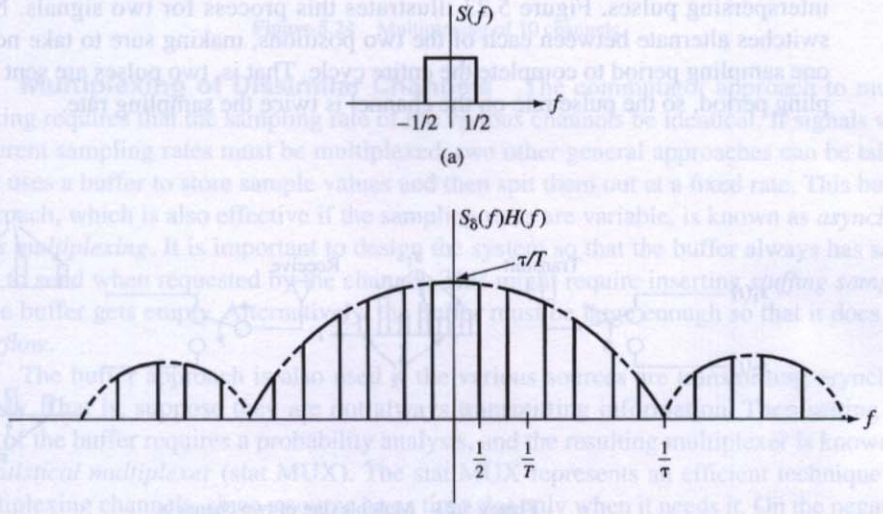


Figure 5.22 Result for Example 5.4.

A significant general observation to make about the transform of a PAM wave is that it occupies all frequencies, from zero to infinity.

5.3.2 Time Division Multiplexing

It would be very impractical to have a separate channel for every signal to be communicated. In telephone communication, this would mean a wire connection for every conversation; if up to 100,000 simultaneous calls between the United States and Europe were anticipated, 100,000 channels (i.e., pairs of wires or satellite channels) would be needed. An even more absurd example is provided by television: If 13 television stations wanted to broadcast simultaneously (forgetting uhf), we would need 13 parallel atmospheric systems—even the science fiction writers would have trouble with that. So we need a way to share a channel among different users.

Signals can be easily separated from each other if they are nonoverlapping in either time or frequency. An example of separation in time occurs in the classroom when students and the professor alternately share the communication channel: When a student talks, the professor stops, and vice versa. An exact dual of this is separation in frequency. Suppose one speaker has unusual vocal cords, and his or her voice occupies frequencies between 1 kHz and 2 kHz. Suppose a second speaker is a baritone, all of whose essential signal components are below 1 kHz. Filters (lowpass and bandpass) can easily separate these two signals even if the two people speak simultaneously. We will deal extensively with separation in frequency beginning with the next chapter. For now, we concentrate on separation in time. Fortunately, over portions of the time axis the PAM waveform is zero, so that separation in time is possible.

Time division multiplexing of signals with identical sampling rates can be viewed as interspersing pulses. Figure 5.23 illustrates this process for two signals. Note that the switches alternate between each of the two positions, making sure to take no longer than one sampling period to complete the entire cycle. That is, two pulses are sent in each sampling period, so the pulse rate on the channel is twice the sampling rate.

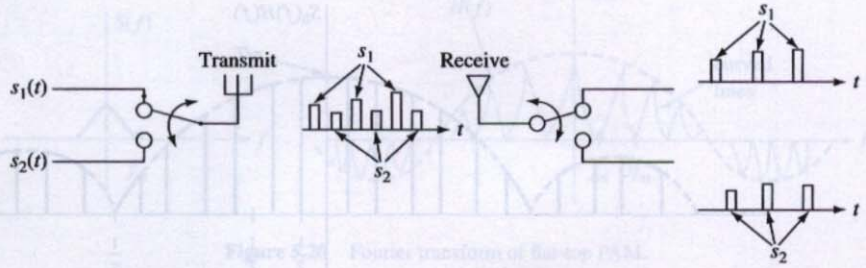


Figure 5.23 Multiplexing of two channels.

Suppose that we now increase the number of channels from 2 to 10. Then the switch becomes a commutator, as shown in Fig. 5.24. The switch must make one complete rotation fast enough so that it arrives at Channel 1 in time for the second sample. The rotation of the receiver switch must be synchronized with that at the transmitter. In practice, this synchronization (known as *frame synchronization*) requires effort. If we knew exactly what was being sent on one of the channels, we could identify its samples at the receiver. Indeed, a common method of synchronization is to sacrifice one of the channels and send a known synchronizing signal in its place. We shall see this in some of the digital transmission systems in later chapters.

The only thing that limits how fast the switch can rotate, and therefore how many channels can be multiplexed, is the fraction of time required for each PAM signal. That fraction is the ratio of the width of each pulse to the spacing between adjacent samples of a single channel. The trade-off design consideration is that the more narrow each pulse, the wider will be the bandwidth of the resulting signal.

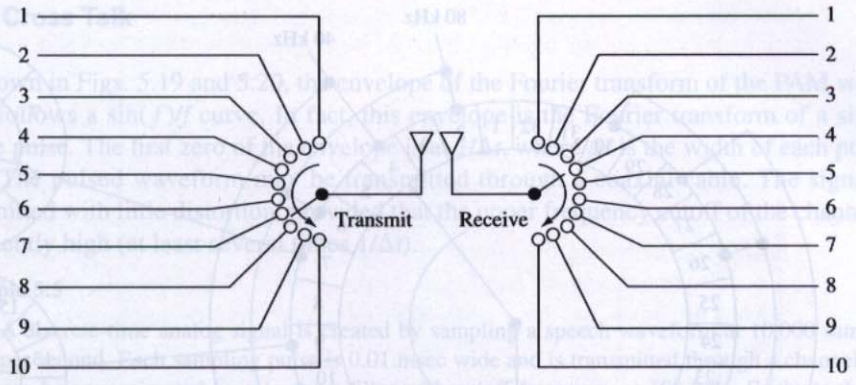


Figure 5.24 Multiplexing of 10 channels.

Multiplexing of Dissimilar Channels The commutator approach to multiplexing requires that the sampling rate of the various channels be identical. If signals with different sampling rates must be multiplexed, two other general approaches can be taken. One uses a buffer to store sample values and then spit them out at a fixed rate. This buffer approach, which is also effective if the sampling rates are variable, is known as *asynchronous multiplexing*. It is important to design the system so that the buffer always has samples to send when requested by the channel. This might require inserting *stuffing samples* if the buffer gets empty. Alternatively, the buffer must be large enough so that it does not *overflow*.

The buffer approach is also used if the various sources are transmitting asynchronously. That is, suppose they are not always transmitting information. Then setting the size of the buffer requires a probability analysis, and the resulting multiplexer is known as a *statistical multiplexer* (stat MUX). The stat MUX represents an efficient technique for multiplexing channels, since a source has a time slot only when it needs it. On the negative side, since individual messages are not being transmitted at a regular rate, the messages must be *tagged* with a user ID. If the channels are synchronous with samples occurring at a regular and continuous rate, the stat MUX is not the best approach.

The second general approach involves *sub-* and *supercommutation*. This requires that all sampling rates be multiples of some basic rate. Meeting such a requirement might necessitate sampling some of the channels at a rate higher than what you would use without multiplexing. For example, if you have two channels with required sampling rates of 8 kHz and 15.5 kHz, then in order to effect that combination, you might choose to sample the faster channel at 16 kHz.

The concept of sub- and supercommutation is quite simple, and we illustrate it with an example. Figure 5.25 shows a *commutator wheel* with 32 slots. Suppose we wish to multiplex the following 44 channels:

1 channel sampled at 80 kHz
1 channel sampled at 40 kHz
18 channels sampled at 10 kHz
8 channels sampled at 1,250 Hz
16 channels sampled at 625 Hz

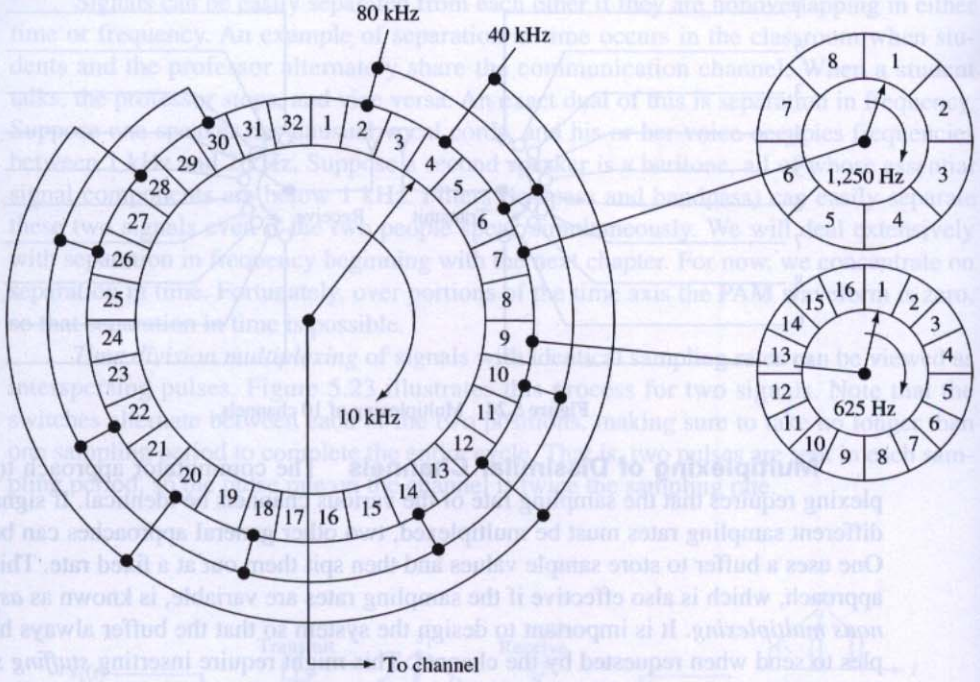


Figure 5.25 Example of sub- and supercommutation.

All of the sampling rates are multiples of 625 Hz. Let us choose the basic rate of the commutator wheel to be 10,000 rotations per second. Therefore, each of the 18 channels that must be sampled at 10 kHz get one slot on the wheel. The channel that must be sampled at 40 kHz needs four equally spaced slots on the wheel, so it is sampled four times during each 0.1-msec rotation of the wheel. Similarly, the 80-kHz channel needs eight equally spaced slots on the wheel. These higher rate channels are multiplexed using *supercommutation*.

The channels sampled at less than 10 kHz need to be sampled only on selected rotations of the wheel. For example, a 1,250-Hz channel needs to be sampled once every 8 rotations of the wheel, while a 625-Hz channel is sampled only once every 16 rotations. We accomplish this using *subcommutation* wheels, as shown in the figure. The eight 1,250-Hz channels are commutated together with a wheel rotating at a rate of 1,250 rotations per second. Each 0.1 msec, one of the channels is connected to a cell on the main commutator wheel. Similarly, the 16 625-Hz channels are commutated with a wheel rotating at 625 rotations per second.

Clearly, a lot of design work had to go into this configuration. We have chosen numbers that work out perfectly. Life is usually not so nice, however, and manipulation is needed to design the commutation system. In some ways, this process resembles finding the lowest common denominator in combining fractions, but of course, it is orders of magnitude more complex than this simple algebraic problem.

5.3.3 Cross Talk

As shown in Figs. 5.19 and 5.20, the envelope of the Fourier transform of the PAM waveform follows a $\sin(f)/f$ curve. In fact, this envelope is the Fourier transform of a single square pulse. The first zero of the envelope is at $1/\Delta t$, where Δt is the width of each pulse.

The pulsed waveform may be transmitted through a coaxial cable. The signal is transmitted with little distortion, provided that the upper frequency cutoff of the channel is sufficiently high (at least several times $1/\Delta t$).

Example 5.5

A discrete-time analog signal is created by sampling a speech waveform at 10,000 samples per second. Each sampling pulse is 0.01 msec wide and is transmitted through a channel that can be approximated by a lowpass filter with cutoff frequency at 100 kHz. Evaluate the effects of channel distortion.

Solution

Since the sampling is occurring at 10,000 samples per second, we will assume that the speech waveform has a maximum frequency below 5 kHz. The reason for the two-to-one ratio was discussed in Section 5.2. The transmitted signal consists of pulses 0.01 msec wide. If one of these pulses forms the input to a lowpass filter with cutoff at 100 kHz, the output of the filter

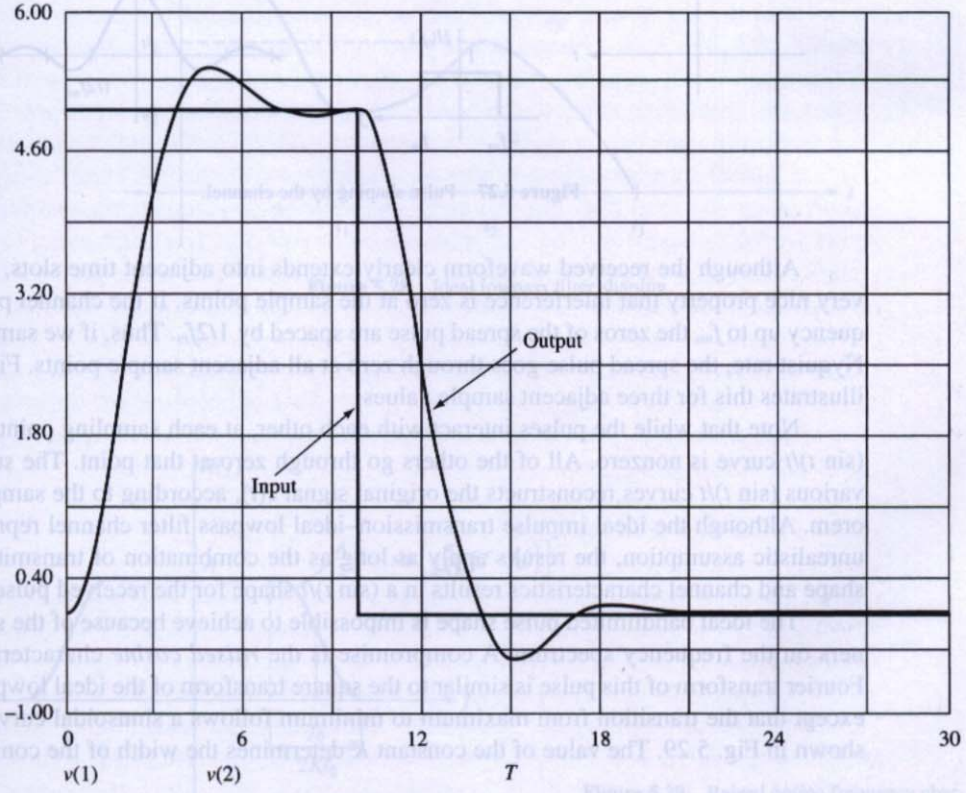


Figure 5.26 Pulse through lowpass filter.

is as shown in Figure 5.26.[1] Thus, although the original pulse may be confined to its assigned time interval, the filtering effects of the channel may widen the pulse to overlap adjacent intervals.

The overlap from one time slot to adjacent time slots is known as *intersymbol interference* or *crosstalk*. The term *crosstalk* also applies to the leakage of signals from one set of wires to an adjacent set of wires, as when multiple wires form part of one cable. We will restrict our discussion to the overlap of time slots in multiplexed systems. There are several ways of reducing the effects of intersymbol interference. Pulse spreading can be decreased by increasing the bandwidth of the system. Unfortunately, this is a luxury that requires a flexibility we don't often have. However, one parameter over which we do have control is the shape of the pulses used to transmit the sample values.

As a start toward the analysis of desirable pulse shapes, suppose we transmit ideal impulse samples. Suppose further that the channel can be modeled as an ideal lowpass filter. This shapes each impulse into a (sin t)/t type of pulse, as shown in Fig. 5.27.

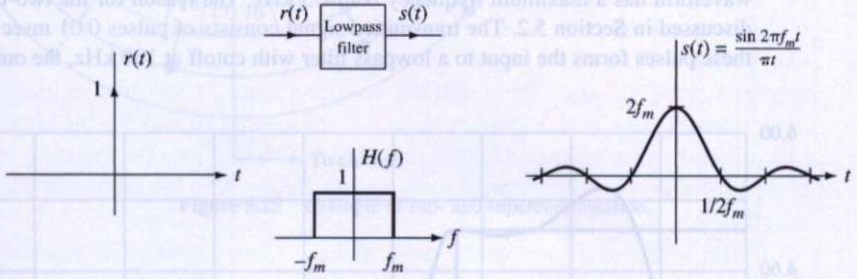


Figure 5.27 Pulse shaping by the channel.

Although the received waveform clearly extends into adjacent time slots, it has the very nice property that interference is zero at the sample points. If the channel passes frequency up to f_m, the zeros of the spread pulse are spaced by $1/2f_m$. Thus, if we sample at the Nyquist rate, the spread pulse goes through zero at all adjacent sample points. Figure 5.28 illustrates this for three adjacent sample values.

Note that while the pulses interact with each other, at each sampling point only one (sin t)/t curve is nonzero. All of the others go through zero at that point. The sum of the various (sin t)/t curves reconstructs the original signal $s(t)$, according to the sampling theorem. Although the ideal impulse transmission–ideal lowpass filter channel represents an unrealistic assumption, the results apply as long as the combination of transmitted pulse shape and channel characteristics results in a (sin t)/t shape for the received pulse.

The ideal bandlimited pulse shape is impossible to achieve because of the sharp corners on the frequency spectrum. A compromise is the *raised cosine* characteristic. The Fourier transform of this pulse is similar to the square transform of the ideal lowpass filter, except that the transition from maximum to minimum follows a sinusoidal curve. This is shown in Fig. 5.29. The value of the constant K determines the width of the constant por-

[1]We have used a circuit simulation program, MICRO-CAP IV, to produce the output waveform. The channel was approximated as a third-order Butterworth lowpass filter with normalized transfer function $1/(s^3 + 2s^2 + 2s + 1)$.

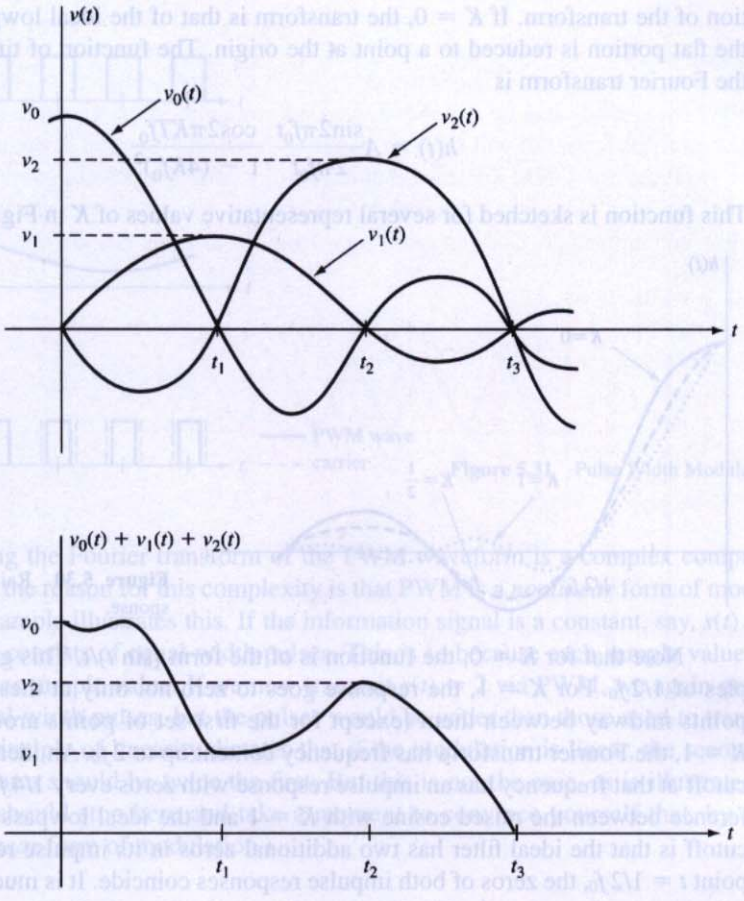


Figure 5.28 Ideal lowpass filter shaping.

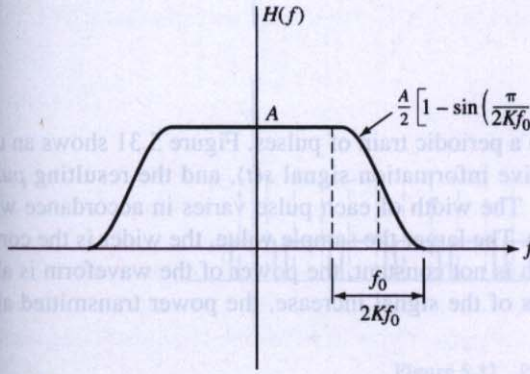


Figure 5.29 Raised cosine frequency characteristic.

tion of the transform. If $K = 0$, the transform is that of the ideal lowpass filter. If $K = 1$, the flat portion is reduced to a point at the origin. The function of time corresponding to the Fourier transform is

$$h(t) = A\frac{\sin 2\pi f_0 t}{2\pi f_0 t}\frac{\cos 2\pi KTf_0}{1-(4Kf_0)^2} \tag{5.14}$$

This function is sketched for several representative values of K in Fig. 5.30.

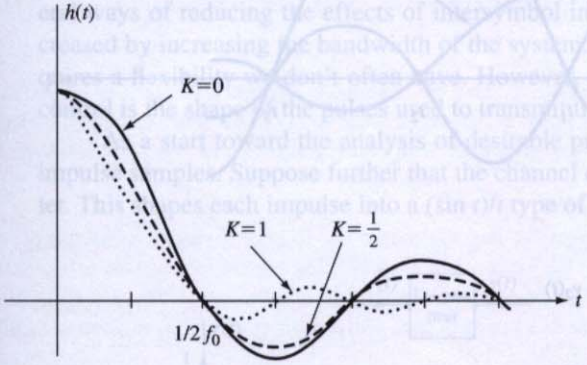


Figure 5.30 Raised cosine impulse response.

Note that for $K = 0$, the function is of the form $(\sin t)/t$. This goes to zero at multiples of $1/2f_0$. For $K = 1$, the response goes to zero not only at these points, but also at points midway between them (except for the first set of points around the origin). For $K = 1$, the Fourier transform has frequency content up to $2f_0$. An ideal lowpass filter with cutoff at that frequency has an impulse response with zeros every $1/4f_0$. Therefore, the difference between the raised cosine with $K = 1$ and the ideal lowpass filter with the same cutoff is that the ideal filter has two additional zeros in its impulse response. Beyond the point $t = 1/2f_0$, the zeros of both impulse responses coincide. It is much easier to build the raised cosine filter than the ideal lowpass filter; indeed, the latter cannot be built in the real world. The price we pay is intersymbol interference between adjacent samples (no interference more than one sample period away). We can compensate for this by using a technique known as *partial response signaling*. In digital communication, this is called *duobinary*. It is a form of controlled intersymbol interference. Since we know the proportion of one sample value that interacts with the next sample point, we can compensate by giving our receiver memory.

5.3.4 Pulse Width Modulation

Let us again start with a signal that is a periodic train of pulses. Figure 5.31 shows an unmodulated pulse train, a representative information signal $s(t)$, and the resulting *pulse width–modulated* (PWM) waveform. The width of each pulse varies in accordance with the instantaneous sample value of $s(t)$. The larger the sample value, the wider is the corresponding pulse. Since the pulse width is not constant, the power of the waveform is also not constant. Thus, as the amplitudes of the signal increase, the power transmitted also increases.

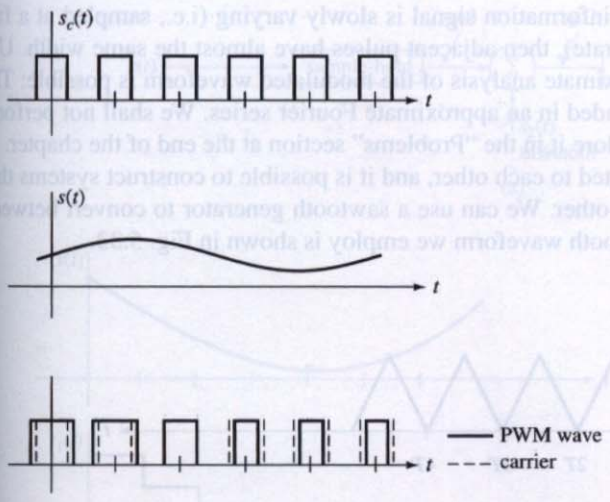


Figure 5.31 Pulse Width Modulation.

Finding the Fourier transform of the PWM waveform is a complex computational task. Part of the reason for this complexity is that PWM is a *nonlinear* form of modulation. A simple example illustrates this. If the information signal is a constant, say, $s(t) = 1$, the PWM wave consists of equal-width pulses. This is so because each sample value is equal to every other sample value. If we now transmit $s(t) = 2$ via PWM, we again get a pulse train of equal-width pulses, but the pulses would be wider than those used to transmit $s(t) = 1$. The principle of linearity dictates that if the modulation is linear, the second modulated waveform should be twice the first. But this is not the case, as is illustrated in Fig. 5.32. (You should stop here and take a moment to convince yourself that, by contrast, PAM is a *linear* form of modulation.)

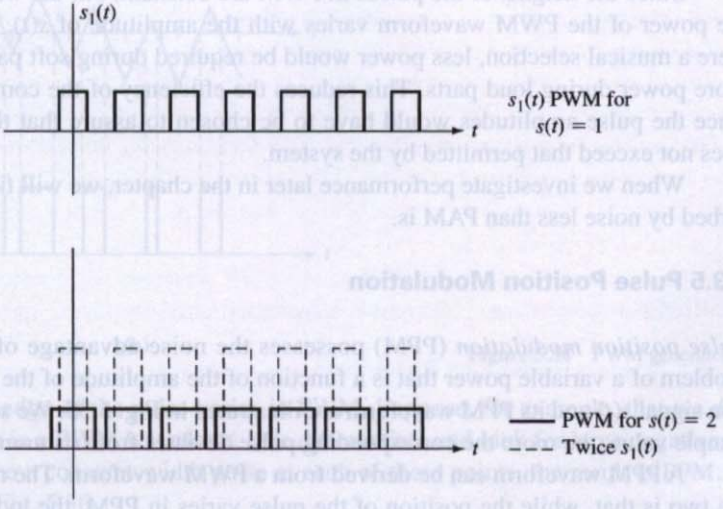


Figure 5.32 PWM is nonlinear.

If one assumes that the information signal is slowly varying (i.e., sampled at a fast rate compared to the Nyquist rate), then adjacent pulses have almost the same width. Under this assumption, an approximate analysis of the modulated waveform is possible: The PWM waveform can be expanded in an approximate Fourier series. We shall not perform the analysis here, but will explore it in the "Problems" section at the end of the chapter.

PAM and PWM are related to each other, and it is possible to construct systems that convert from one form to the other. We can use a sawtooth generator to convert between time and amplitude. The sawtooth waveform we employ is shown in Fig. 5.33.

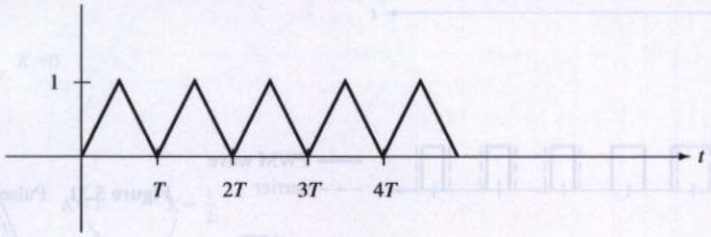


Figure 5.33 Sawtooth waveform for PWM-to-PAM conversion.

The conversion process is illustrated in Fig. 5.34. Figure 5.34(a) shows a block diagram of the generator, and Fig. 5.34(b) shows typical waveforms.

We start with an information signal $s(t)$. This is put through a sample-and-hold circuit to yield $s_1(t)$. The sawtooth is shifted down by one unit in order to form $s_2(t)$. The sum of $s_1(t)$ and $s_2(t)$ is $s_3(t)$. The times for which $s_3(t)$ is positive represent intervals whose widths are proportional to the original sample values. We need only put the shifting sawtooth into a comparator with output of 1 for positive input and 0 for negative input. This results in $s_4(t)$, the PWM waveform. The range of pulse widths can be adjusted by scaling the original function of time. Our illustration assumes that the original $s(t)$ was normalized to lie between 0 and 1.

Since the heights of the pulses in PWM are constant, but the widths depend on $s(t)$, the power of the PWM waveform varies with the amplitude of $s(t)$. For example, if $s(t)$ were a musical selection, less power would be required during soft parts of the music and more power during loud parts. This reduces the efficiency of the communication system, since the pulse amplitudes would have to be chosen to assure that the maximum power does not exceed that permitted by the system.

When we investigate performance later in the chapter, we will find that PWM is disturbed by noise less than PAM is.

5.3.5 Pulse Position Modulation

Pulse position modulation (PPM) possesses the noise advantage of PWM without the problem of a variable power that is a function of the amplitude of the signal. An information signal $s(t)$ and its PPM waveform are illustrated in Fig. 5.35. We see that the larger the sample value, the more the corresponding pulse deviates from its unmodulated position.

A PPM waveform can be derived from a PWM waveform. The relationship between the two is that, while the position of the pulse varies in PPM, the location of the leading

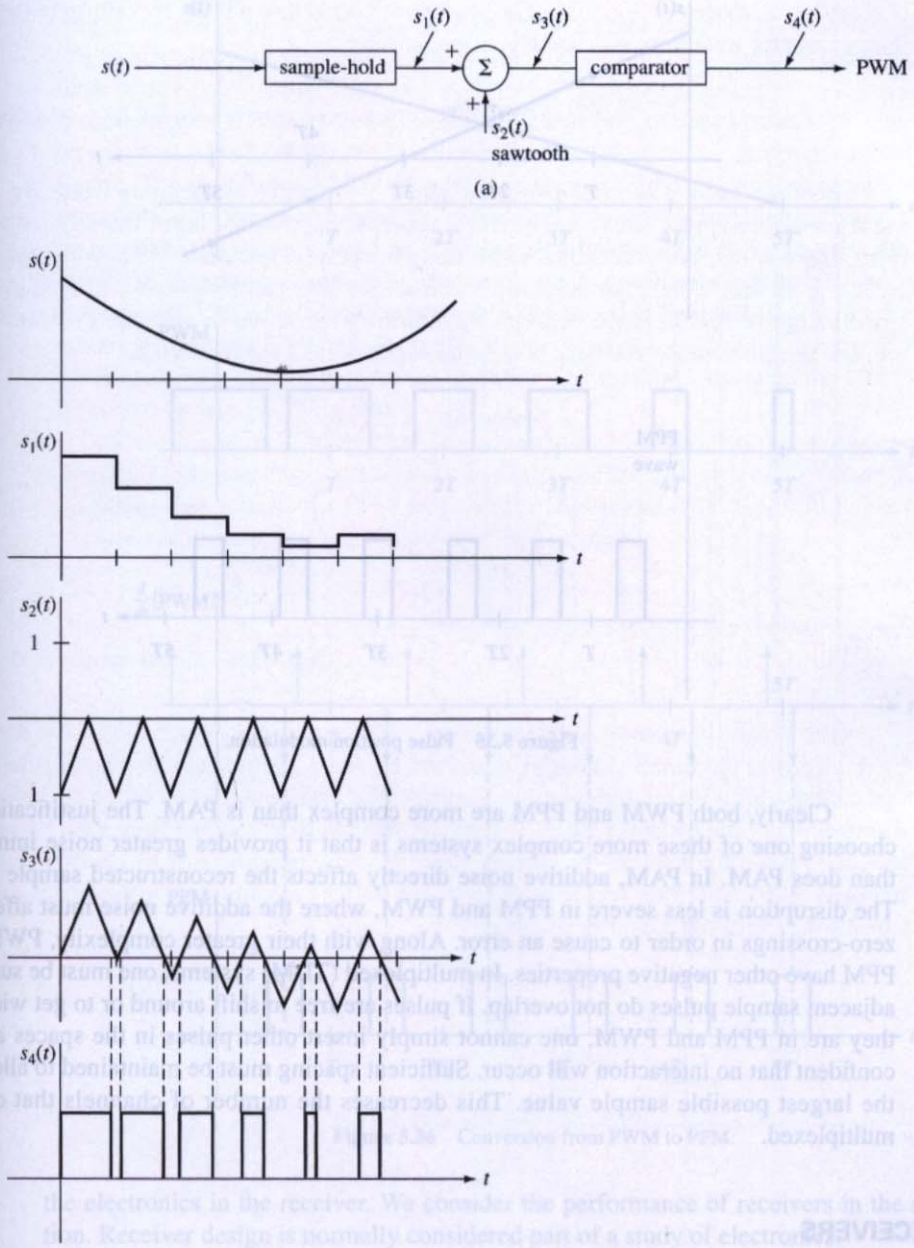


Figure 5.34 PWM generator.

(or trailing) edge of the pulse varies in PWM. Suppose, for example, that we detect each trailing edge in a PWM waveform. (We differentiate and look for large negative pulses.) If we now place a constant-width pulse at each of these points, the result is PPM. This is illustrated in Fig. 5.36.

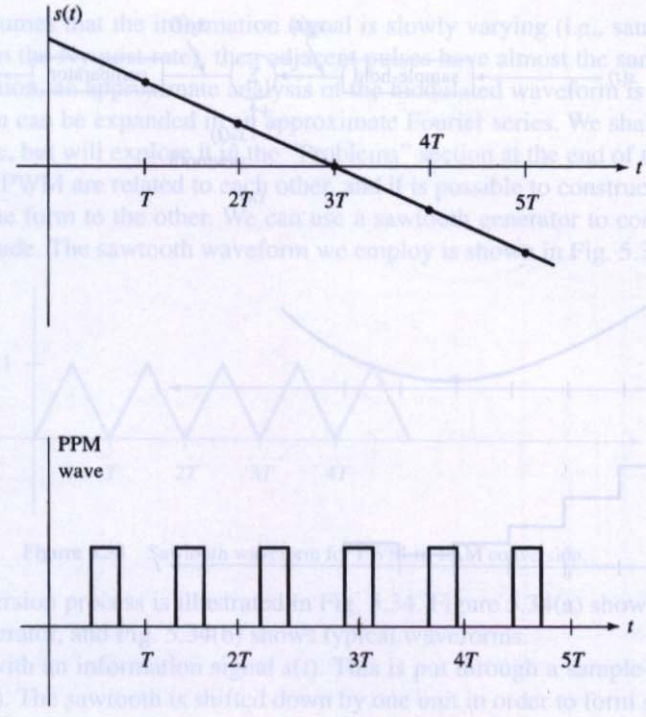


Figure 5.35 Pulse position modulation.

Clearly, both PWM and PPM are more complex than is PAM. The justification for choosing one of these more complex systems is that it provides greater noise immunity than does PAM. In PAM, additive noise directly affects the reconstructed sample value. The disruption is less severe in PPM and PWM, where the additive noise must affect the zero-crossings in order to cause an error. Along with their greater complexity, PWM and PPM have other negative properties. In multiplexed (TDM) systems, one must be sure that adjacent sample pulses do not overlap. If pulses are free to shift around or to get wider, as they are in PPM and PWM, one cannot simply insert other pulses in the spaces and be confident that no interaction will occur. Sufficient spacing must be maintained to allow for the largest possible sample value. This decreases the number of channels that can be multiplexed.

5.4 RECEIVERS

5.4.1 Analog Baseband Reception

Analog baseband reception consists of filtering the received signal, amplifying it, and feeding the result into a transducer (e.g., a speaker). In the case of audio, the receiver is simply an audio amplifier. In designing receivers, one needs to be concerned with filter characteristics and noise. Noise is added in the channel, and additional noise is added by

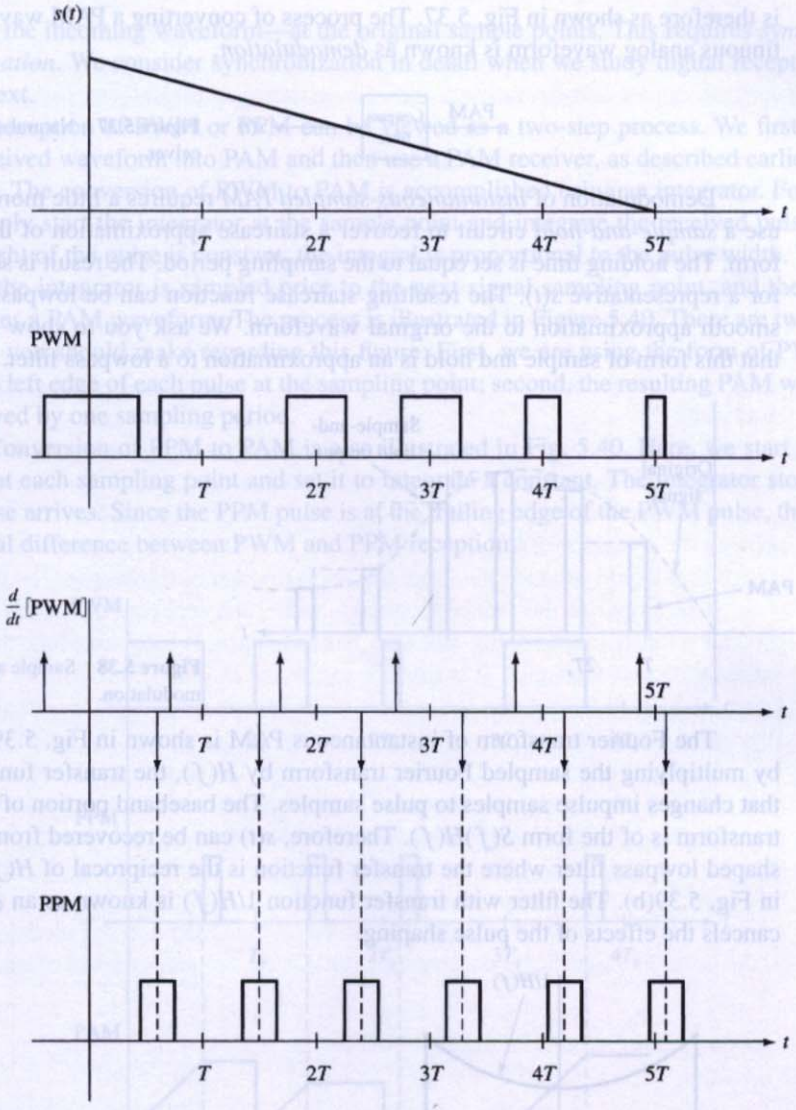


Figure 5.36 Conversion from PWM to PPM.

the electronics in the receiver. We consider the performance of receivers in the next section. Receiver design is normally considered part of a study of electronics.

5.4.2 Discrete Baseband Reception

Reconstruction of the original signal from *natural-sampled PAM* follows directly from the sampling theorem. Indeed, natural-sampled PAM is the type of signal processing we encountered in our first proof of the sampling theorem. Recovery of the original analog signal from its sampled version requires a lowpass filter. The natural-sampled PAM receiver

is therefore as shown in Fig. 5.37. The process of converting a PAM waveform to the continuous analog waveform is known as *demodulation*.

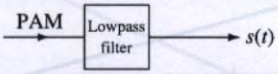


Figure 5.37 Natural-sampled PAM receiver.

Demodulation of *instantaneous-sampled PAM* requires a little more work. We could use a *sample-and-hold* circuit to recover a staircase approximation of the original waveform. The holding time is set equal to the sampling period. The result is shown in Fig. 5.38 for a representative $s(t)$. The resulting staircase function can be lowpass filtered to get a smooth approximation to the original waveform. We ask you to show in Problem 5.4.1 that this form of sample and hold is an approximation to a lowpass filter.

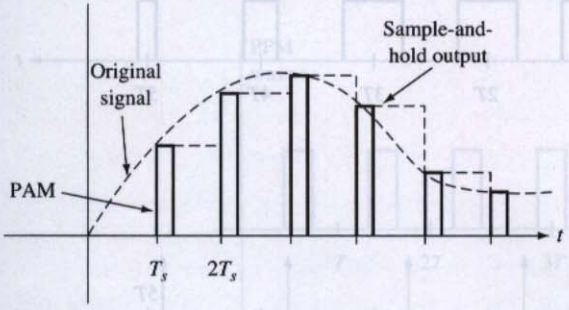
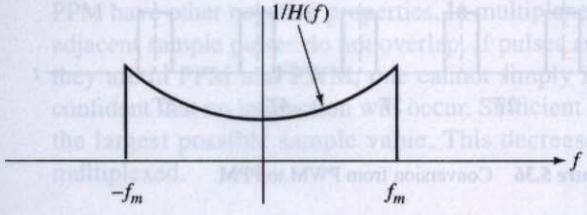
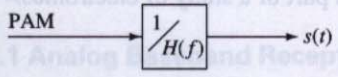


Figure 5.38 Sample and hold for PAM demodulation.

The Fourier transform of instantaneous PAM is shown in Fig. 5.39(a). We derive it by multiplying the sampled Fourier transform by $H(f)$, the transfer function of the filter that changes impulse samples to pulse samples. The baseband portion of the PAM Fourier transform is of the form $S(f)H(f)$. Therefore, $s(t)$ can be recovered from $s_m(t)$ by using a shaped lowpass filter where the transfer function is the reciprocal of $H(f)$. This is shown in Fig. 5.39(b). The filter with transfer function $1/H(f)$ is known as an *equalizer,* since it cancels the effects of the pulse shaping.



(a) PAM transform



(b) Equalizer

Figure 5.39 Flat-top PAM demodulator.

The equalizer and lowpass filter demodulators do not require that the receiver recover timing information. On the other hand, the sample-and-hold demodulator *does* require such timing information at the receiver. That is, the receiver must "know" when to

sample the incoming waveform—at the original sample points. This requires *symbol synchronization*. We consider synchronization in detail when we study digital reception later in the text.

Reception of PWM or PPM can be viewed as a two-step process. We first convert the received waveform into PAM and then use a PAM receiver, as described earlier in this section. The conversion of PWM to PAM is accomplished using an integrator. For PWM, we simply start the integrator at the sample point and integrate the received pulse. Since the height of the pulse is constant, the integral is proportional to the pulse width. The output of the integrator is sampled prior to the next signal sampling point, and the sample generates a PAM waveform. The process is illustrated in Figure 5.40. There are two observations you should make regarding this figure: First, we are using the form of PWM that sets the left edge of each pulse at the sampling point; second, the resulting PAM waveform is delayed by one sampling period.

Conversion of PPM to PAM is also illustrated in Fig. 5.40. Here, we start the integrator at each sampling point and set it to integrate a constant. The integrator stops when the pulse arrives. Since the PPM pulse is at the trailing edge of the PWM pulse, there is no essential difference between PWM and PPM reception.

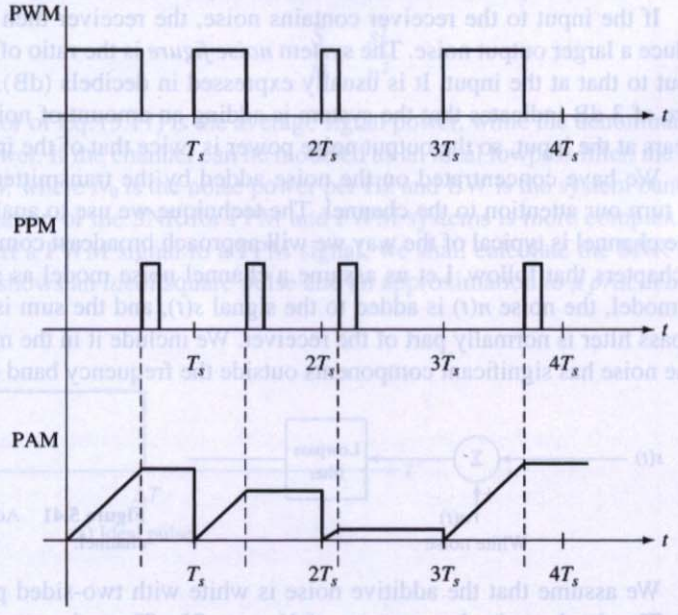


Figure 5.40 Conversion of PWM and PPM to PAM.

5.5 PERFORMANCE

We will learn to design a variety of communication systems. Performance evaluation needs to apply to a wide range of system inputs. In analog communication, we normally wish the output to be as close to the input waveform as possible. The more common measure of such closeness is the S/N power ratio, since the human ear is sensitive to this quan-

tity. In general, the ear can hear additive disturbances if the ratio of S/N power is below a certain threshold.

In digital communication, the normal measure of performance is the rate at which bit errors occur.

5.5.1 Analog Baseband

The output SNR of a baseband analog receiver depends on the input SNR, the filtering characteristic of the receiver, and the noise added by the electronics in the receiver.

The SNR at the input to the receiver depends on the characteristics of the channel and of the noise that intrudes during transmission. We normally consider this additive noise to be white Gaussian, so the total power of that noise is proportional to the system bandwidth.

The noise added by the receiver is characterized by the *noise figure. Thermal noise* is produced by the random motion of electrons in a medium.

If we have a system with a number of noise-generating devices within it, we often refer to the system *noise temperature* T_e, in °K. This is the temperature of a single noise source that would produce the same total noise power at the output.

If the input to the receiver contains noise, the receiver then adds its own noise to produce a larger output noise. The system *noise figure* is the ratio of the noise power at the output to that at the input. It is usually expressed in decibels (dB). For example, a noise figure of 3 dB indicates that the system is adding an amount of noise equal to that which appears at the input, so the output noise power is twice that of the input.

We have concentrated on the noise added by the transmitter and the receiver. We now turn our attention to the channel. The technique we use to analyze the additive noise in the channel is typical of the way we will approach broadcast communication systems in the chapters that follow. Let us assume a channel noise model as shown in Fig. 5.41. In this model, the noise $n(t)$ is added to the signal $s(t)$, and the sum is lowpass filtered. The lowpass filter is normally part of the receiver. We include it in the model because, without it, the noise has significant components outside the frequency band of interest.

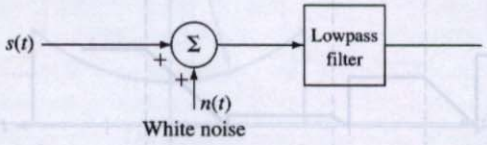


Figure 5.41 Additive noise in baseband channel.

We assume that the additive noise is white with two-sided power spectral density $N_0/2$. That is, the noise has a power of N_0 watts/Hz. The noise power at the output of the filter is then $N_0 f_m$ watts, and

$$\text{SNR} = \frac{P_s}{N_0 f_m} \tag{5.15}$$

where P_s is the power of the signal.

The performance of a baseband analog receiver also depends on nonlinearities in the electronics. This is expressed in measures such as dynamic range and harmonic distortion. The term *dynamic range* usually refers to the ratio (in decibels) of the strongest to the

weakest signal that a receiver can process without noise or distortion exceeding acceptable limits. Although this sounds like a simple concept, application to practical transmission is quite complex. For example, the behavior of a receiver when a single sinusoid forms the input may be quite different from that when the input is a complex sum of many signal components. Keep in mind that we are discussing nonlinear effects, and the actual dynamic range of the receiver may depend upon characteristics of the input signal.

Harmonic distortion is normally measured by setting the receiver input to be a single sinusoid. Nonlinearities in the receiver change this sinusoid to a periodic function with harmonics. The ratio of the power of the harmonics to the power of the fundamental is a measure of harmonic distortion.

5.5.2 Discrete Baseband

The SNR in a PAM signal depends on the form of the receiver. If the receiver simply samples the received waveform at periodic points in time, the sample values are

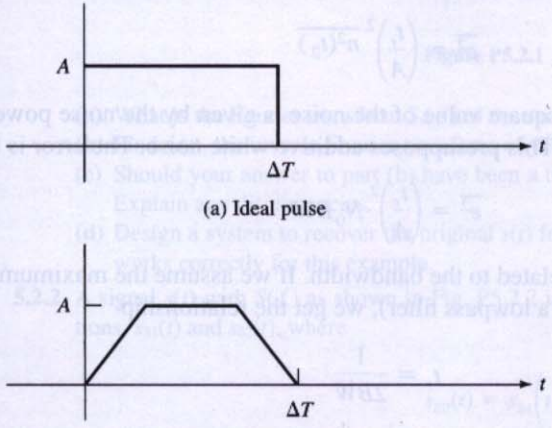
$$r(nT_s) = s(nT_s) + n(nT_s) \tag{5.16}$$

where $n(t)$ is the additive noise. The SNR is then

$$\frac{S}{N} = \frac{\overline{s^2}}{\overline{n^2}} \tag{5.17}$$

The numerator of Eq. (5.17) is the average signal power, while the denominator is the average noise power. If the channel can be modeled as an ideal lowpass filter, the noise power is simply N_0BW, where N_0 is the noise power per Hz and BW is the system bandwidth.

Calculation of the SNR for PPM and PWM systems is more complex. Since we can easily convert a PWM signal to a PPM signal, we shall calculate the SNR only for PPM. Figure 5.42 shows an *ideal* square pulse and an approximation to a *practical* square pulse.



(a) Ideal pulse

(b) Approximation to practical pulse

Figure 5.42 Ideal and practical square pulse.

The job of the PPM receiver is to locate the trailing edge of the latter pulse. One way to do this is with a comparator or threshold detector. A threshold is set, and when the signal breaks through it, we assume that we have located the trailing edge. The threshold

value would normally be set at the midpoint, $A/2$, of the pulse amplitude, and this value is known as the *slicing level*. In the case of the *ideal* square pulse, as long as the additive noise waveform never exceeds the slicing level in magnitude, the location of the trailing edge will not be affected.

If we now add noise to the practical approximation to the square pulse, we have the situation shown in Fig. 5.43, where we have focused upon the trailing edge of the pulse.

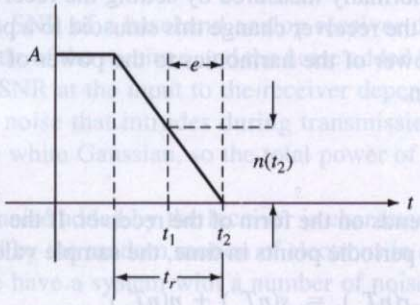


Figure 5.43 Noise affecting location of trailing edge.

Similar triangles can be used to derive the relationship

$$\frac{e}{n(t_2)} = \frac{t_r}{A} \tag{5.18}$$

where A is the amplitude of the pulse, t_r is the pulse rise time, and $n(t_2)$ is the additive noise at the time the perturbed signal crosses the slicing level. We solve for the timing error to obtain

$$e = \left(\frac{t_r}{A}\right) n(t_2) \tag{5.19}$$

The mean square value of the error is then

$$\overline{e^2} = \left(\frac{t_r}{A}\right)^2 \overline{n^2(t_2)} \tag{5.20}$$

Let us assume that the mean square value of the noise is given by the noise power per Hz multiplied by the bandwidth. This presupposes additive white noise. The error is then

$$\overline{e^2} = \left(\frac{t_r}{A}\right)^2 N_0 BW \tag{5.21}$$

The pulse rise time is related to the bandwidth. If we assume the maximum possible slope (a square pulse through a lowpass filter), we get the relationship

$$t_r = \frac{1}{2BW} \tag{5.22}$$

Equation (5.22) is related to the sampling theorem. We also would like to relate the error to the pulse energy instead of the amplitude, since many systems are energy limited. The pulse energy is

$$E_p \approx A^2 \Delta T \tag{5.23}$$

The approximation improves as the rise time decreases. Combining Eqs. (5.21), (5.22), and (5.23) yields

$$\overline{e^2} = \frac{\Delta T}{4BWE_p} N_0 \tag{5.24}$$

Equation (5.24) is the desired result. It shows that the mean square timing error is inversely proportional to the system bandwidth. To convert the PPM waveform to an analog signal, we change the pulse location to a pulse amplitude and then take the ratio of the mean square value of the signal samples to the mean square value of the error. The noise power is therefore related to the mean square timing error of Eq. (5.24). We explore this relationship in the "Problems" section at the back of this chapter.

PROBLEMS

5.2.1 You are given the function

$$s(t) = \frac{\sin t}{t}$$

This function is sampled by the train of impulses, $s_\delta(t)$ as shown in Fig. P5.2.1.

$$s_s(t) = s(t)s_\delta(t)$$

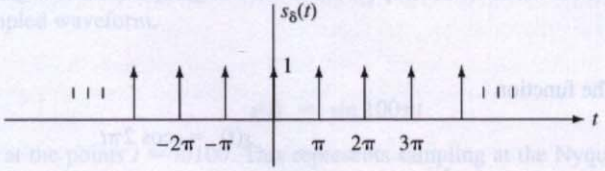


Figure P5.2.1

(a) What is the Fourier transform $S_s(f)$ of the sampled function?
(b) Find $s_s(t)$, the inverse Fourier transform of your answer to part (a).
(c) Should your answer to part (b) have been a train of impulses? Did it turn out that way? Explain any discrepancies.
(d) Design a system to recover the original $s(t)$ from $s_s(t)$, and demonstrate that your system works correctly for this example.

5.2.2 A signal $s(t)$ with $S(f)$ as shown in Fig. P5.2.2 is sampled by two different sampling functions, $s_{\delta 1}(t)$ and $s_{\delta 2}(t)$, where

$$s_{\delta 2}(t) = s_{\delta 1}\left(t - \frac{T}{2}\right)$$

and

$$T = \frac{1}{f_m}$$

Therefore, each of the two sampling waveforms is at one-half of the Nyquist minimum sampling frequency. Find the Fourier transforms of the sampled waveforms, $s_{s1}(t)$ and $s_{s2}(t)$.

Now consider $s(t)$ to be sampled by $s_{\delta 3}(t)$, a train of impulses spaced $T/2$ apart (i.e., at the Nyquist rate). This new sampling function is the sum $s_{\delta 1}(t) + s_{\delta 2}(t)$. Show that the Fourier transform of $s_{s3}(t)$ is equal to the sum of the transforms of $s_{s2}(t)$ and $s_{s2}(t)$. That is, show that, although the transforms of the original two sampled functions contain aliasing error, this error is not present in the sum. [*Hint:* You will have to keep track of phases.]

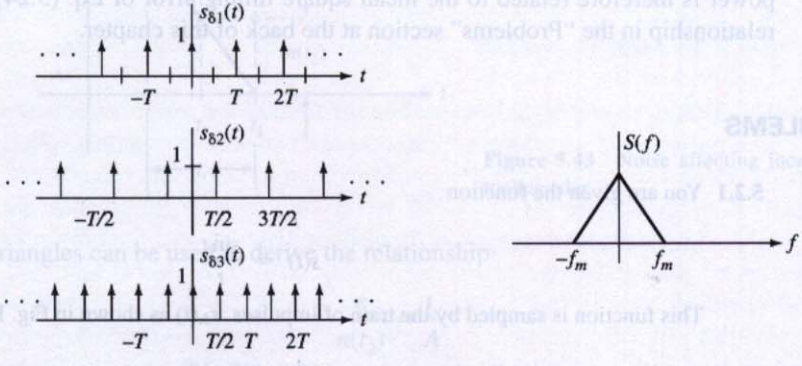


Figure P5.2.2

5.2.3 The function

$$s(t) = \cos 2\pi t$$

is sampled every $\frac{3}{4}$ second. Evaluate the aliasing error.

5.2.4 The signal

$$s(t) = \frac{\sin 2\pi t}{\pi t}$$

is sampled at 1.1 times the Nyquist rate. The sampling is performed for t between -1 and $+1$ second. The signal is reconstructed using a lowpass filter. Find the truncation error.

5.2.5 The function

$$s(t) = \cos 2\pi t$$

is sampled at a rate of 2.5 samples/sec for t between 0 and 10 seconds. The signal is reconstructed using a lowpass filter. Find the difference between the original and reconstructed waveforms at the points $t = 4.9$, $t = 5$, and $t = 5.1$ sec.

5.2.6 You are given a low-frequency bandlimited signal $s(t)$. This signal is multiplied by the pulse train $s_c(t)$, as shown in Fig. P5.2.6. Find the Fourier transform of the product $s(t)s_c(t)$. What restrictions must be imposed so that $s(t)$ can be uniquely recovered from the product waveform?

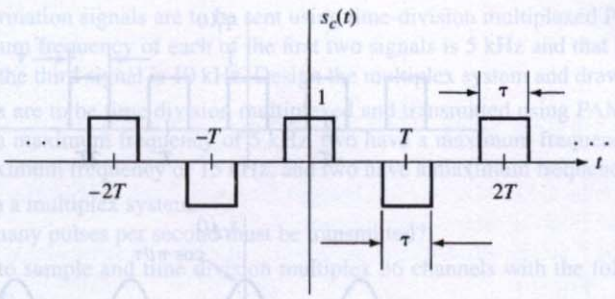


Figure P5.2.6

5.2.7 A signal is given by

$$s(t) = \frac{\sin 5\pi t}{\pi t} + \frac{\sin 10\pi t}{\pi t}$$

This signal is to be sampled with a periodic pulse train consisting of narrow pulses.

(a) Find the Nyquist sampling rate.

(b) Assuming that the sampling is done at the Nyquist rate, sketch the Fourier transform of the sampled waveform.

5.2.8 The signal

$$s(t) = \sin 100\pi t$$

is sampled at the points $t = n/100$. This represents sampling at the Nyquist rate of 100 Hz. However, all of the sample values will be zero, and the original wave cannot be reconstructed. Explain the reason for this situation.

5.2.9 A signal is of the form

$$s(t) = \sin \pi t + 3\sin 3\pi t$$

It is sampled at 1.5 times the Nyquist rate and transmitted using PAM. The pulse waveform, $s_c(t)$, is a periodic train of triangular pulses, as shown in Fig. 5.21. Find the Fourier transform of the modulated waveform.

5.2.10 Derive the dual of the time-sampling theorem—that is, a Fourier transform of a time-limited signal $s(t)$ is completely known from its sample values. Find the minimum frequency spacing between samples to permit reconstruction of the Fourier transform.

5.3.1 An information signal is of the form

$$s(t) = \frac{\sin \pi t}{\pi t}$$

Find the Fourier transform of the waveform that results if each of the two carrier waveforms shown in Fig. P5.3.1 is pulse amplitude modulated with this signal.

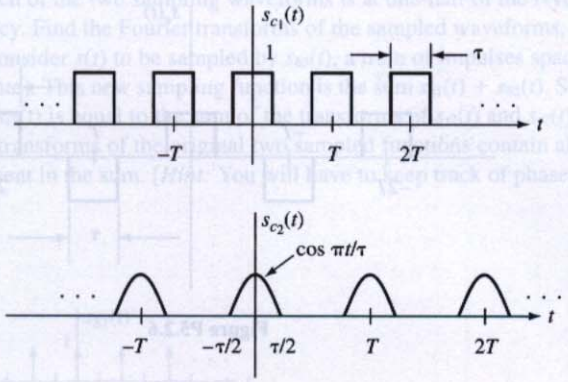


Figure P5.3.1

5.3.2 The signal

$$s(t) = \cos 2\pi t$$

is sampled every 0.4 sec and sent using natural-sampled PAM with pulse widths of 0.1 sec. The channel can be modeled as an ideal lowpass filter with a cutoff at 10 Hz. Find the received waveform. Also, find the reconstructed waveform after the receiver uses a lowpass filter to recover the original signal $s(t)$.

5.3.3 Consider a two-channel TDM PAM system where both channels are used to transmit the same signal $s(t)$ with Fourier transform $S(f)$, as shown in Fig. P5.3.3. The system samples $s(t)$ at the minimum rate. Find the Fourier transform of the TDM waveform, and compare it to the Fourier transform of a single-channel PAM system used to transmit $s(t)$.

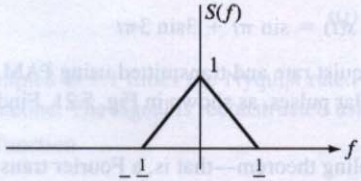


Figure P5.3.3

5.3.4 Three asynchronous sources transmit PAM waveforms to a buffer multiplexer. Each of the sources transmits at a pulse rate that is Gaussian distributed with a mean value of 1,000 pulses/sec and a variance of 9. The channel is capable of transmitting 3,000 pulses per second. How large must the buffer be such that the probability of overload is less than 1%?

5.3.5 Three information signals are to be sent using time-division multiplexed PAM. Suppose that the maximum frequency of each of the first two signals is 5 kHz and that the maximum frequency of the third signal is 10 kHz. Design the multiplex system and draw a block diagram.

5.3.6 Ten signals are to be time division multiplexed and transmitted using PAM. Four of the signals have a maximum frequency of 5 kHz, two have a maximum frequency of 10 kHz, two have a maximum frequency of 15 kHz, and two have a maximum frequency of 20 kHz.

(a) Design a multiplex system.

(b) How many pulses per second must be transmitted?

5.3.7 You wish to sample and time division multiplex 36 channels with the following maximum frequencies:

One channel has a maximum of 10 kHz.

Three channels have a maximum of 5 kHz.

Eight channels have a maximum of 2.5 kHz.

Eight channels have a maximum of 300 Hz.

Sixteen channels have a maximum of 150 Hz.

Design a system using sub- and supercommutation. Make any reasonable approximations.

5.3.8 A discrete-time analog signal is created by sampling a speech waveform at 10,000 samples per second. Each sampling pulse is 0.01 msec wide and is transmitted through a channel that can be approximated by a lowpass filter with cutoff frequency at 50 kHz. Evaluate the effects of channel distortion. Compare your answer to that of Example 5.5.

5.3.9 A PWM system multiplexes three signals derived from waveforms with the same maximum frequency f_m. Sample values are normalized (to lie between zero and unity), and the width w of each pulse is related to the normalized sample value $s_i(nT_s)$ by

$$w(nT_s) = 1 + s_i(nT_s) \ \mu\text{sec}$$

(a) What is the maximum frequency f_m of the baseband signals that would permit this multiplexing to take place?

(b) What is the minimum bandwidth of the channel?

5.4.1 Show that a sample-and-hold circuit is an approximation to a lowpass filter, provided that the sampling is performed at the Nyquist rate or higher. [*Hint:* You may wish to find the step response of the sample-and-hold circuit and compare it to the step response of a lowpass filter.]

5.4.2 The system shown in Fig. P5.4.2 is similar to a sample-and-hold circuit, where the input can be considered to be an impulse-sampled version of the original waveform.

(a) Find the impulse response of this system.

(b) Find the transfer function, and compare it to the transfer function of a lowpass filter.

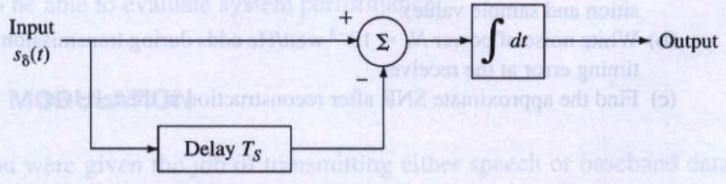


Figure P5.4.2

5.4.3 Two signals,

$$s_1(t) = \cos 2\pi t$$

$$s_2(t) = \cos \pi t + 2\cos 2\pi t$$

are sampled every 0.4 sec and are sent using multiplexed natural-sampled PAM. The channel can be modeled as an ideal lowpass filter with a cutoff at 10 Hz.

(a) Find the received waveform.

(b) Find the reconstructed waveforms after the receiver uses a lowpass filter and demultiplexer.

(c) Repeat part (b), with $s_1(t)$ changed to $\cos 1.9\pi t$.

5.4.4 You wish to investigate the use of a first-order lowpass filter (e.g., an *RC* circuit) to recover the original time signal from a PWM waveform. You can assume that each PWM pulse has its leading edge at the sample point. Analyze this system, and comment on how well it acts as a demodulator.

5.5.1 Consider the system shown in Fig. P5.5.1. We wish to compare $y(t)$ to $x(t)$ in order to evaluate the sample-and-hold circuit as a PAM demodulator. The comparison between $y(t)$ and $x(t)$ is performed by defining an error

$$e = \frac{1}{T}\int_0^T [y(t) - x(t)]^2 dt$$

Find the value of this error term.

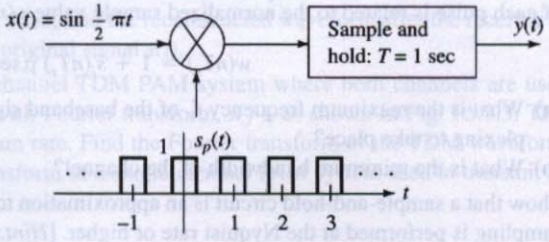


Figure P5.5.1

5.5.2 A sinusoidal signal

$$s(t) = \sin 2\pi t$$

is sampled every 0.4 sec and transmitted using PPM.

(a) Design the PPM system (i.e., choose a pulse width and a relationship between pulse position and sample value).

(b) White noise of power $N_0 = 10^{-3}$ watt/Hz adds during transmission. Find the mean square timing error at the receiver.

(c) Find the approximate SNR after reconstruction at the receiver.

6 Amplitude Modulation

6.0 PREVIEW

What We Will Cover and Why You Should Care

This is the first of two chapters dealing with modulation techniques for analog communication. You will learn the basic concepts of modulation and examine the motivation for using various modulation schemes. After reading the chapter, you will:

- Understand amplitude modulation and the difference between suppressed and transmitted carrier modulation
- Know how to construct modulators
- Know how to construct demodulators
- Know how standard broadcast AM radio works
- Understand the various types of AM stereo
- Know how to perform video transmission (e.g., TV)
- Possess the necessary tools to evaluate and compare the performance of systems.

Necessary Background

The earlier portions of this chapter require that you have an understanding of Fourier transforms. You will need to know systems analysis in order to be in a position to design modulators and demodulators. A working knowledge of random processes and probability is needed to be able to evaluate system performance.

6.1 CONCEPT OF MODULATION

Suppose you were given the job of transmitting either speech or baseband data through a channel. The first question you should ask yourself is whether the signals must be modified before injecting them into the channel. If the answer is no, your job is very simple: You must simply decide how to couple the signal into the channel (i.e., interface the two).

For many channels, the answer will be yes, and the signal will have to be modified. The Fourier transform of a typical speech waveform is sketched in Figure 6.1.

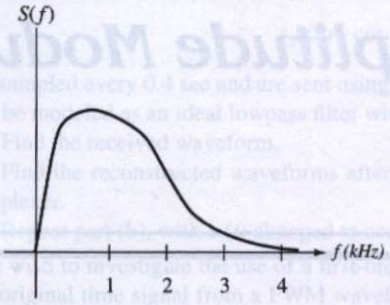


Figure 6.1 Fourier transform of baseband signals.

In the case of short-range transmission, as in the local loop of a telephone circuit or in the path between the pre-amp and amplifier or between the amplifier and speakers of a sound system, these baseband (low-frequency) signals are sent through wires. For longer distances, it is sometimes difficult to use wires, since they require *rights of way*. Additionally, since transmission is *point to point*, one must specify the location of *every* terminal. In the case of television, the wire would have to terminate in the home of every prospective viewer (as in cable television). Mobile communication by wire is almost impossible. (We say *almost* because some missiles actually trail a wire behind them that unwinds as does a fishing line—but this is the exception.) For all of these reasons, *broadcast* communication has been a popular form of transmission.

Suppose we take an audio signal and attempt to transmit it through the air. Let us choose a typical audio frequency of 1 kHz. The wavelength of a 1-kHz signal in air is approximately 300 km (about 180 miles). A quarter-wavelength antenna would then have to be 75 km (45 miles) long, and erecting such antennas in backyards of homes would be a bit impractical! But even if we were willing to erect them, we would still be left with two very serious problems. The first is related to the characteristics of air at audio frequencies: While propagation does occur at frequencies below 10 kHz, these frequencies are not efficiently transmitted through air. Even more serious is the second problem: interference. Often, it is desirable to transmit more than one analog signal at a time. For example, many local radio stations transmit broadcasts simultaneously. If they used quarter-wavelength antennas, they would each have antennas 75 km long on top of their studios (or on mountaintops), and they would pollute the air with many audio signals. The listener would erect an antenna 75 km high and receive a weighted sum of all of the signals (depending on relative distances and antenna patterns from the different transmitting antennas to the receiving antenna). Since the only information the receiver would have about the signals is that they would all be bandlimited to the same upper cutoff frequency, there would be absolutely no way of separating the signal from one station from those from all of the others.[1]

6.2

Given the preceding scenario, it is desirable to modify a low-frequency signal before sending it from one point to another. An added bonus arises if the modified signal is less susceptible to noise than is the original signal.

[1]Walk into a crowded, noisy room, and try to distinguish one conversation from all of the others. Then record the sounds in the room, and try again to distinguish the sounds, this time by listening to the recording. Ask yourself why there is a difference.

The most common method of accomplishing the modification is to use the low-frequency signal to modulate (i.e., modify the parameters of) another, higher frequency signal. Most commonly, this other signal is a pure sinusoid.

We start, then, with a pure sinusoid $s_c(t)$ called the *carrier* waveform. It is given this name because it is used to *carry* the information signal from the transmitter to the receiver. Mathematically,

$$s_c(t) = A \cos(2\pi f_c t + \theta) \tag{6.1}$$

If f_c is properly chosen, this carrier waveform can be efficiently transmitted. For example, suppose you were told that frequencies in the range between 1 MHz and 3 MHz propagate in a mode that allows them to be reliably sent over distances up to about 200 km. If you chose the frequency f_c to be in this range, then the pure sinusoidal carrier would transmit efficiently. The wavelength of transmission in the range of 1 MHz to 3 MHz is on the order of 100 meters, and antennas of reasonable length can be used.

We now ask the question whether the preceding pure sinusoidal carrier waveform can somehow be altered in a way that (a) the altered waveform still propagates efficiently and (b) the information we wish to send is somehow superimposed on the new waveform in a way that it can be recovered at the receiver. In other words, we are asking whether there is some way that the sinusoid can *carry* the information along. The answer is yes, as we now illustrate.

The right-hand side of Eq. (6.1) contains three parameters that may be varied: the amplitude A, the frequency f_c, and the phase θ. Using the information signal to vary A, f_c, or θ leads to *amplitude modulation, frequency modulation,* and *phase modulation,* respectively.

We will show that efficient transmission is achieved for each of these three cases. We will also show that if more than one signal is simultaneously propagated through the channel, separation of the signals at the receiver is possible. In addition, we will find it critical to illustrate a third property: The information signal $s(t)$ must be uniquely recoverable from the received modulated waveform; it would not be of much use to modify a carrier waveform for efficient transmission and station separability if we could not reproduce $s(t)$ accurately at the receiver.

This chapter concentrates on a thorough treatment of amplitude modulation. Parallel treatments of frequency and phase modulation follow in the next chapter.

6.2 DOUBLE-SIDEBAND SUPPRESSED CARRIER

If we modulate the amplitude of the carrier of Eq. (6.1), the following modulated waveform results:

$$s_m(t) = A(t)\cos(2\pi f_c t + \theta) \tag{6.2}$$

The frequency f_c and the phase θ are constant. The amplitude $A(t)$ varies somehow in accordance with the baseband signal $s(t)$–the signal we want carried through the channel.

We simplify the expression by assuming that $\theta = 0$. This will not affect any of the basic results, since the angle actually corresponds to a time shift of $\theta/2\pi f_c$. A time shift is not considered distortion in a communication system.

If somebody asked you how to vary $A(t)$ in accordance with $s(t)$, the simplest answer you could suggest would be to make $A(t)$ equal to $s(t)$. This would yield a modulated signal of the form

$$s_m(t) = s(t)\cos 2\pi f_c t \tag{6.3}$$

Such a signal is given the name *double-sideband suppressed carrier* (DSBSC) *amplitude modulation* for reasons that will soon become clear.

This simple equating of $A(t)$ with $s(t)$ does indeed satisfy the criteria demanded of a communication system. The easiest way to illustrate this fact is to express $s_m(t)$ in the frequency domain, that is, to find its Fourier transform.

Suppose that we let $S(f)$ be the Fourier transform of $s(t)$. We require nothing more of $S(f)$ than that it be the Fourier transform of a baseband signal. That is, $S(f)$ must equal zero for frequencies above some cutoff frequency f_m. (The subscript m stands for *maximum*.) Figure 6.2 gives a representative sketch of $S(f)$. We do not mean to imply that $S(f)$ must be of the shape shown; the sketch is meant only to indicate the transform of a general low-frequency bandlimited signal.

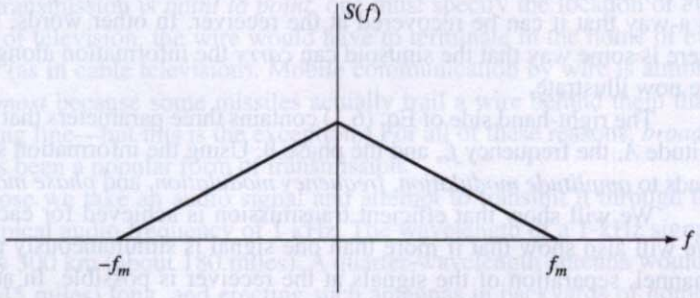


Figure 6.2 General form of baseband $S(f)$.

The modulation theorem is used to find $S_m(f)$:

$$S_m(f) = \mathcal{F}[s(t)\cos 2\pi f_c t] = \frac{1}{2}[S(f + f_c) + S(f - f_c)] \tag{6.4}$$

This transform is sketched as Figure 6.3. Note that modulation of a carrier with $s(t)$ has shifted the frequencies of $s(t)$ both up and down by the frequency of the carrier. This is

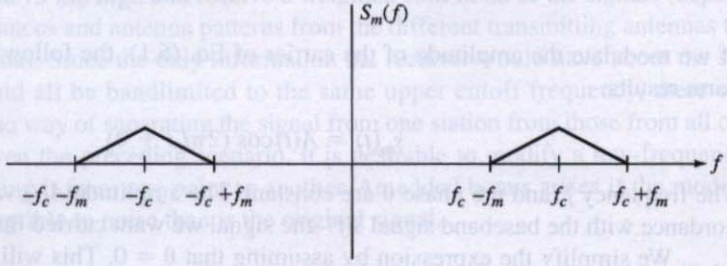


Figure 6.3 $S_m(f)$, the transform of $s_m(t)$.

analogous to the trigonometric result that multiplication of a sinusoid by another sinusoid results in sum and difference frequencies. That is,

$$\cos A \cos B = \frac{1}{2}\cos(A + B) + \frac{1}{2}\cos(A - B) \tag{6.5}$$

If cos A is replaced by $s(t)$, where $s(t)$ contains a continuum of frequencies between 0 and f_m, the trigonometric identity can be applied term by term to yield a result containing all sums and differences of the frequencies.

Figure 6.3 indicates that the modulated waveform $s_m(t)$ contains components with frequencies between $f_c - f_m$ and $f_c + f_m$. As long as signals in this range of frequencies transmit efficiently and an antenna of reasonable length can be constructed, we have solved the first of the two problems. Let us plug in some typical audio numbers. Let f_m be 5 kHz and f_c be 1 MHz. Then the range of frequencies occupied by the modulated waveform is from 995,000 to 1,005,000 Hz.

The second objective is separation of the signals. We see that if one information signal modulates a sinusoid of frequency f_{c1} and another information signal modulates a sinusoid of frequency f_{c2}, the Fourier transforms of the two modulated carriers do not overlap in frequency, provided that f_{c1} and f_{c2} are separated by at least $2f_m$. This is illustrated in Figure 6.4. Since the signals are "stacked" in frequency, we refer to this situation as *frequency*

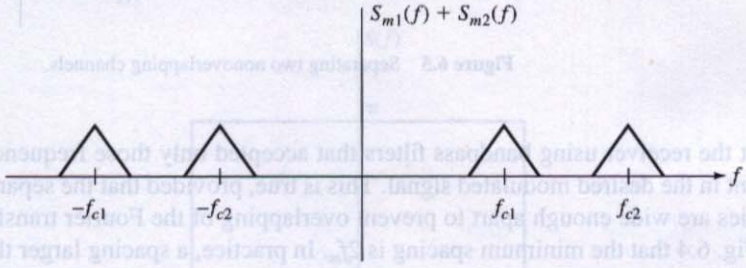


Figure 6.4 Fourier transform of two AM signals.

division multiplexing (FDM). It is the exact dual of time division multiplexing (TDM), which we introduced in Chapter 5.

If the frequencies of the two modulated waveforms are not too widely separated, both signals can even share the same antenna. That is, although the optimum antenna length is not the same for both channels, the total bandwidth can be made relatively small compared to the carrier frequency. In practice, the antenna is usable over a *range* of frequencies rather than just being effective at a single frequency. If this were not true, radio broadcasting would not exist.

As an example, you don't have to readjust the length of your car antenna whenever you tune across the AM dial. The effectiveness of the antenna does not vary greatly from one frequency limit to the other. Instructions accompanying early car antennas suggested that their length be shortened to about 75 centimeters when changing from AM to FM. (Don't try doing this if you have the type of antenna that is sandwiched within the windshield.) Modern receivers have sufficient sensitivity that such tuning is no longer necessary, even with frequency changes of two orders of magnitude.

If signals are nonoverlapping in time, gates or switches can be used to effect their separation. For AM, the signals are nonoverlapping in frequency, and they can be separated from each other by means of frequency gates (bandpass filters). Thus, a system such as that shown in Fig. 6.5 could be used to separate the two modulated carriers of Fig. 6.4 from each other.

The extension of this system to more than two channels should be obvious. Even if many modulated signals were transmitted over the same channel, they could be separated

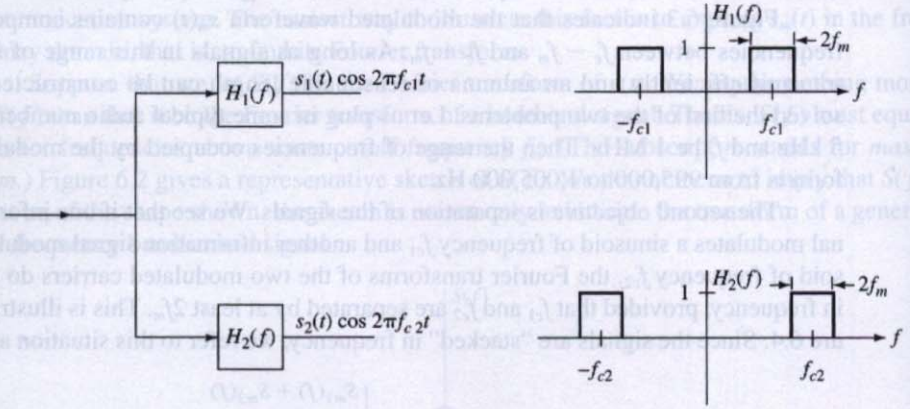


Figure 6.5 Separating two nonoverlapping channels.

at the receiver using bandpass filters that accepted only those frequencies that were present in the desired modulated signal. This is true, provided that the separate carrier frequencies are wide enough apart to prevent overlapping of the Fourier transforms. We see from Fig. 6.4 that the minimum spacing is $2f_m$. In practice, a spacing larger than this is desirable for two reasons: First, even though we may view the information signal as limited to frequencies below f_m, no matter how sharply we lowpass filter it, the signal still has some components above f_m. Second, if the minimum spacing is used, the bandpass filters that separate out the desired channel must be perfect, with flat response in the passband and an infinite roll-off.

Example 6.1

An information signal is of the form

$$s(t) = \frac{\sin 2\pi t}{t}$$

The signal amplitude modulates a carrier of frequency 10 Hz. Sketch the AM waveform and its Fourier transform.

Solution: The AM waveform is given by the equation

$$s_m(t) = \frac{\sin 2\pi t}{t} \cos 20\pi t$$

This function is sketched in Fig. 6.6.

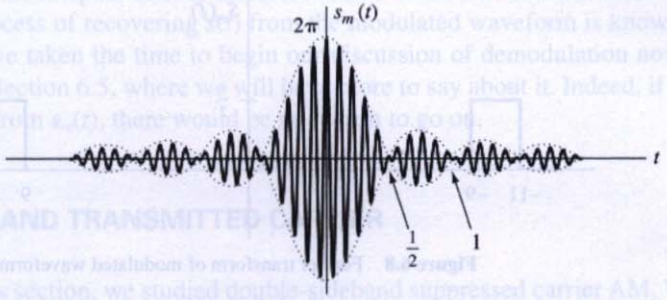


Figure 6.6 AM waveform for Example 6.1.

We note that when the carrier, $\cos 20\pi t$, is equal to 1, $s_m(t) = s(t)$, and when the carrier is equal to -1, $s_m(t) = -s(t)$. In sketching the AM waveform, we start by drawing $s(t)$ and its mirror image, $-s(t)$, as a guide. The AM waveform periodically touches each of these curves and varies smoothly between the periodic points. In this manner, we develop the sketch of the waveform. In most practical situations, the carrier frequency is much higher than that illustrated in this example. In fact, it is so high that if you observed $s_m(t)$ on an oscilloscope, you would not be able to see the back-and-forth oscillations unless you greatly expanded the time axis. Instead, you would see the $s(t)$ and $-s(t)$ outlines and what looks like shading between them.

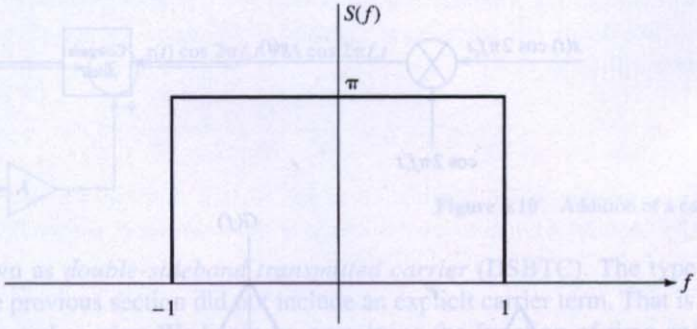


Figure 6.7 Fourier transform of $s(t)$ for Example 6.1.

The Fourier transform of the information signal $s(t)$ is shown in Fig. 6.7. It is found in the table in Appendix II.

The transform of the modulated waveform is given by the following equation, where we have applied the modulation theorem:

$$S_m(f) = \frac{S(f - 10) + S(f + 10)}{2}$$

This is shown in Fig. 6.8.

We have indicated that an AM wave of the type discussed could be transmitted efficiently and that more than one signal could share the channel. A critical property that must still be addressed is whether the information signal, $s(t)$, can be uniquely recovered from the AM waveform.

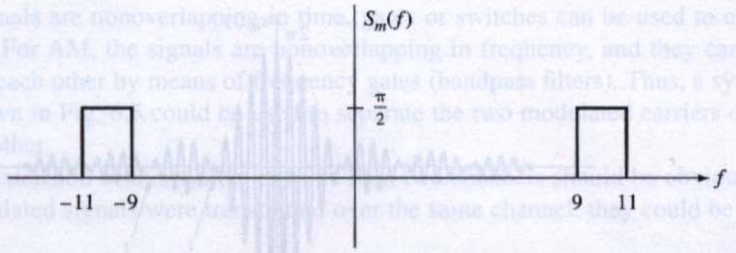


Figure 6.8 Fourier transform of modulated waveform.

Since $S_m(f)$ was derived from $S(f)$ by shifting all of the frequency components of $s(t)$ by f_c, we should be able to recover $s(t)$ from $s_m(t)$ by shifting the frequencies again by the same amount, but this time in the opposite direction.

The *modulation theorem* states that multiplication of a function of time by a sinusoid shifts the Fourier transform of the function both up and down in frequency. Thus, if we *remultiply* $s_m(t)$ by a sinusoid at the carrier frequency, the Fourier transform shifts *back down* to its low-frequency baseband position. The multiplication also shifts the transform up to a position centered about $2f_c$, but this part can easily be rejected using a lowpass filter. The process is illustrated in Fig. 6.9.

The recovery of $s(t)$ is described by the following equations:

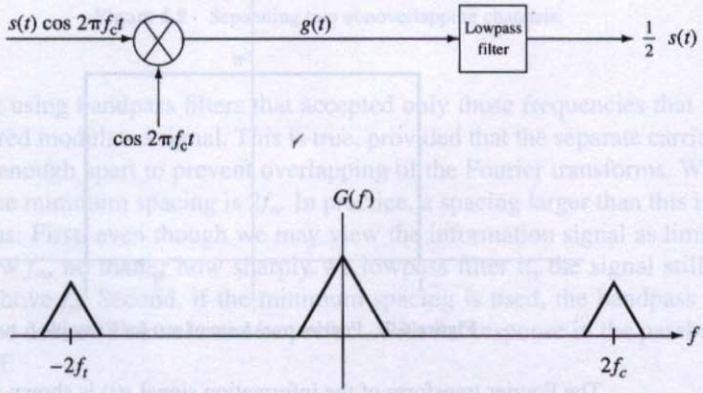


Figure 6.9 Recovery of $s(t)$ from $s_m(t)$.

$$\begin{aligned} s_m(t)\cos 2\pi f_c t &= [s(t)\cos 2\pi f_c t]\cos 2\pi f_c t \\ &= s(t)\cos^2 2\pi f_c t \\ &= \frac{s(t) + s(t)\cos 4\pi f_c t}{2} \end{aligned} \tag{6.6}$$

In Eq. (6.6), we have used the trigonometric identity

$$\cos^2(A) = \frac{1}{2} + \frac{1}{2}\cos(2A) \tag{6.7}$$

The output of the lowpass filter is therefore $s(t)/2$, which is an undistorted version of $s(t)$.

This process of recovering $s(t)$ from the modulated waveform is known as *demodulation*. We have taken the time to begin our discussion of demodulation now, rather than waiting until Section 6.5, where we will have more to say about it. Indeed, if $s(t)$ could not be recovered from $s_m(t)$, there would be no reason to go on.

6.3 DOUBLE-SIDEBAND TRANSMITTED CARRIER

In the previous section, we studied double-sideband suppressed carrier AM. We found that the waveform resulting from the multiplication of the information signal with a carrier sinusoid possesses desirable properties. In particular, the modulation process shifts frequencies from a band around dc to a band around the carrier frequency. This permits efficient transmission and also allows simultaneous transmission of more than one signal.

We now explore a modification of AM in which we add a portion of the pure sinusoidal carrier to the modulated waveform. We will see in Section 6.5 that this addition greatly simplifies the demodulation process.

Figure 6.10 shows the addition of a pure sinusoidal carrier to the double-sideband suppressed carrier waveform. The resulting waveform is

$$s_m(t) = s(t)\cos 2\pi f_c t + A\cos 2\pi f_c t \tag{6.8}$$

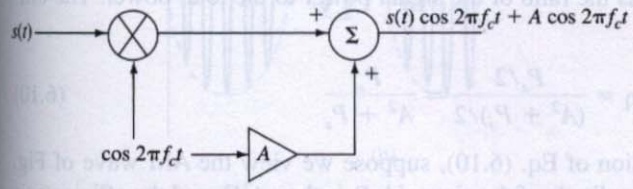


Figure 6.10 Addition of a carrier term.

This is known as *double-sideband transmitted carrier* (DSBTC). The type of AM discussed in the previous section did not include an explicit carrier term. That is why it is labeled *suppressed carrier*. We begin by examining the function of time and its Fourier transform.

The Fourier transform of transmitted carrier AM is the sum of the Fourier transform of suppressed carrier AM and the Fourier transform of the pure carrier. The transform of the carrier is a pair of impulses at $\pm f_c$. The complete transform of the AM wave is therefore as shown in Fig. 6.11.

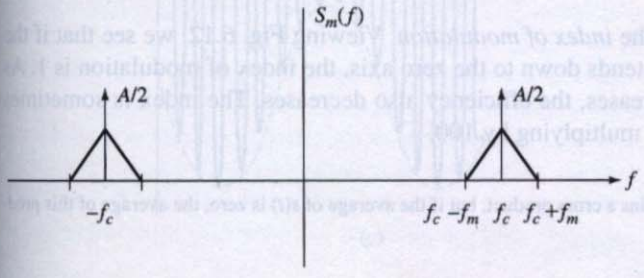


Figure 6.11 Fourier transform of AM transmitted carrier.

The function of time can be sketched if we first combine terms in Eq. (6.8). Doing so, we can rewrite the waveform as

$$s_m(t) = [A + s(t)]\cos 2\pi f_c t \tag{6.9}$$

This function is sketched in the same manner as that used to draw the suppressed carrier waveform. We first draw the outlines at $[A + s(t)]$ and $-[A + s(t)]$. The AM waveform periodically touches these two curves. We then fill in with a smooth, oscillating waveform. This is illustrated for a sinusoidal $s(t)$ [e.g., someone whistling into a microphone] in Figure 6.12.

Figure 6.12(a) shows the sinusoidal $s(t)$, Fig. 6.12(b) shows the AM waveform for a value of A less than the amplitude of $s(t)$, and Fig. 6.12(c) shows the waveform where A is greater than the amplitude of $s(t)$.

Efficiency

We ask you to accept for now the fact that the addition of the carrier makes demodulation easier. The price we pay is in efficiency: A portion of the transmitted power is used to send a pure sinusoid that does not carry any useful information about the signal.

We see from Eq. (6.8) that the carrier power is the power of $A\cos 2\pi f_c t$, or $A^2/2$ watts. The power of the signal portion is the power of $s(t)\cos 2\pi f_c t$, which is the average of $s^2(t)$ divided by 2. The average of $s^2(t)$ is simply the power of $s(t)$, or P_s. Therefore, the signal power is $P_s/2$. The total transmitted power[2] is the sum of this and $A^2/2$.

We define *efficiency*, η, as the ratio of the signal power to the total power. The efficiency is then given by

$$\eta = \frac{P_s/2}{(A^2 + P_s)/2} = \frac{P_s}{A^2 + P_s} \tag{6.10}$$

As an example of the application of Eq. (6.10), suppose we view the AM wave of Fig. 6.12(c) and set A equal to the amplitude of the sinusoid. P_s is then $A^2/2$, and the efficiency is

$$\eta = \frac{A^2/2}{A^2 + A^2/2} = 33\% \tag{6.11}$$

The efficiency depends on the size of the modulated term compared to the size of the pure sinusoidal carrier term. We define a dimensionless quantity m as the ratio of the maximum amplitude of the modulated term to the amplitude of the carrier. That is,

$$m = \frac{\max |s(t)|}{A} \tag{6.12}$$

The quantity m is known as the *index of modulation*. Viewing Fig. 6.12, we see that if the envelope of the waveform extends down to the zero axis, the index of modulation is 1. As the index of modulation decreases, the efficiency also decreases. The index is sometimes expressed as a percentage by multiplying by 100.

[2]The square of the sum contains a cross product, but if the average of $s(t)$ is zero, the average of this product is also zero.

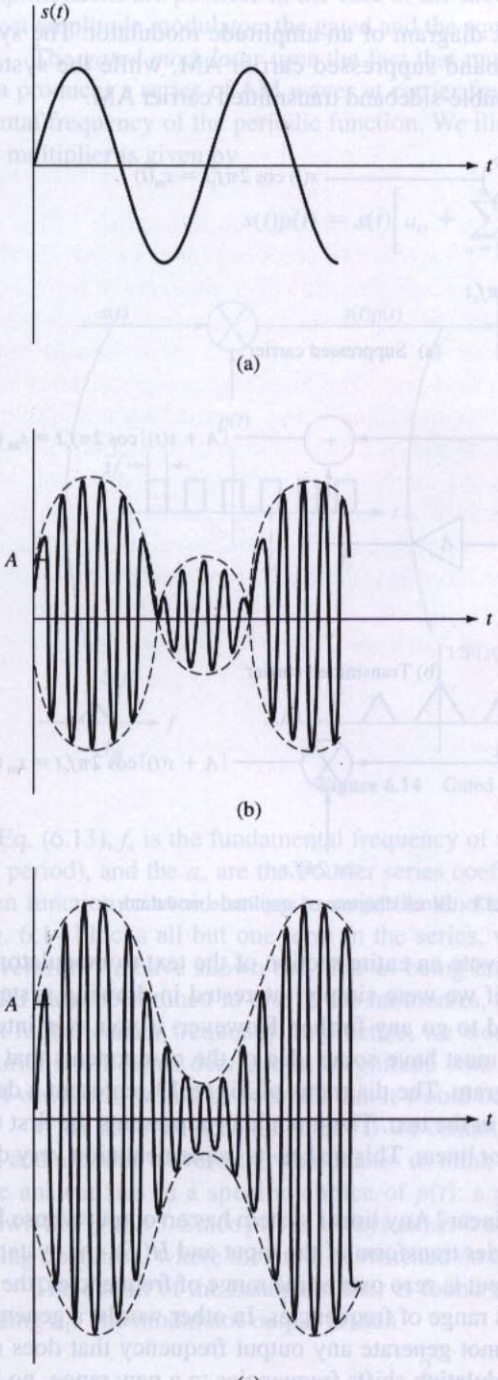
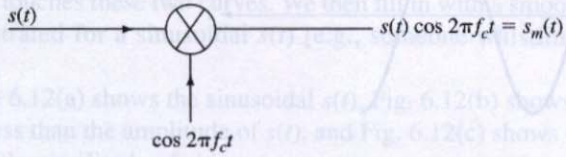


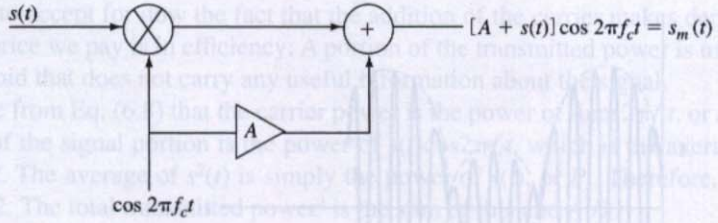
Figure 6.12 AM transmitted carrier waveform.

6.4 MODULATORS

Figure 6.13 presents the block diagram of an amplitude modulator. The system of Fig. 6.13(a) produces double-sideband suppressed carrier AM, while the systems of Figs. 6.13(b) and 6.13(c) produce double-sideband transmitted carrier AM.



(a) Suppressed carrier



(b) Transmitted carrier

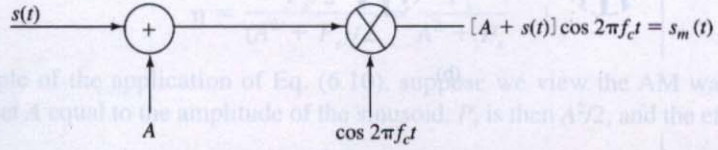


Figure 6.13 Block diagram of amplitude modulator.

You may ask why we devote an entire section of the text to modulators if Fig. 6.13 tells the whole story. Indeed, if we were simply interested in drawing system block diagrams, there would be no need to go any further. However, if you ever intend to implement any system design, you must have some idea of the components that go into each block in the system block diagram. The diagrams of Fig. 6.13 represent a departure from the systems considered earlier in the text. The modulator represents the first time we have come across a system that is *not* linear. This makes its implementation very different from that of linear filters.

Why is modulation not linear? Any linear system has an output whose Fourier transform is the product of the Fourier transform of the input and $H(f)$, the system function. If the Fourier transform of the input is zero over some range of frequencies, the output transform must also be zero for this range of frequencies. In other words, a general property of linear systems is that they cannot generate any output frequency that does not appear in the input. Since amplitude modulation shifts frequencies to a new range, no linear system can perform such an operation.

The synthesis of a system that is not linear is, in general, complicated. Fortunately, simplifications are possible in the case of the modulator. We begin with two classes of indirect amplitude modulator: the gated and the square law modulators.

The *gated modulator* uses the fact that multiplication of $s(t)$ by any periodic function produces a series of AM waves at carrier frequencies that are multiples of the fundamental frequency of the periodic function. We illustrate this in Figure 6.14. The output of the multiplier is given by

$$s(t)p(t) = s(t)\left[a_0 + \sum_{n=1}^{\infty} a_n \cos(2\pi n f_c t)\right] \tag{6.13}$$

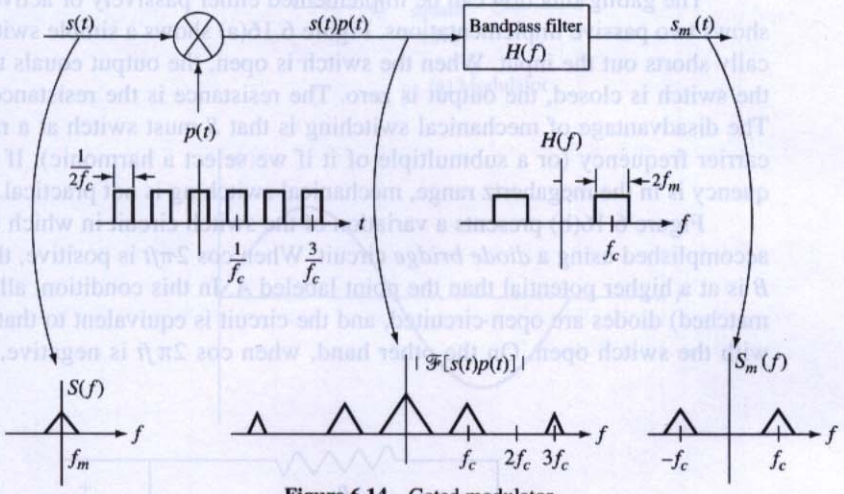


Figure 6.14 Gated modulator.

In Eq. (6.13), f_c is the fundamental frequency of the periodic waveform (the reciprocal of the period), and the a_n are the Fourier series coefficients. We have assumed that $p(t)$ is an even function to avoid having to write the sine terms in the series. The bandpass filter of Fig. 6.14 blocks all but one term in the series, with the result that the output is an AM waveform. We have shown the filter as being tuned to the fundamental frequency, but it could have been tuned to one of the harmonics, thereby resulting in an AM waveform at that higher carrier frequency. In practice, we would favor the lower harmonics, since the Fourier coefficients decrease in magnitude with increasing n. At some point, the output AM waveform would be so small that it would be lost in the circuit noise.

What have we accomplished? If we cannot build a multiplier to take the product of $s(t)$ and a cosine waveform, what makes us think we can build the multiplier of Fig. 6.14? The answer lies in a specific choice of $p(t)$: a periodic pulse train gating function, as shown in Fig. 6.15. Since $p(t)$ is always either 0 or 1, the multiplication can be viewed as a gating operation, where the input is switched on and off.

The output of the bandpass filter is found by expanding $p(t)$ in a Fourier series and finding a_1. The modulator output is then

$$s_m(t) = \frac{2}{\pi} s(t) \cos 2\pi f_c t \tag{6.14}$$

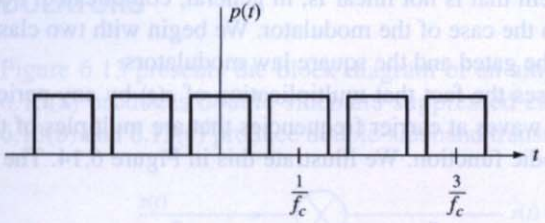
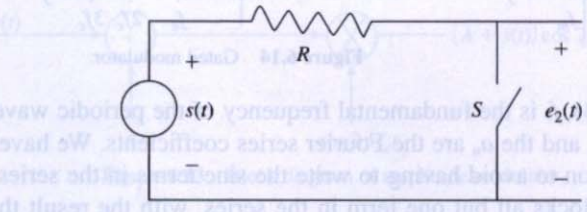


Figure 6.15 Gating with a pulse train.

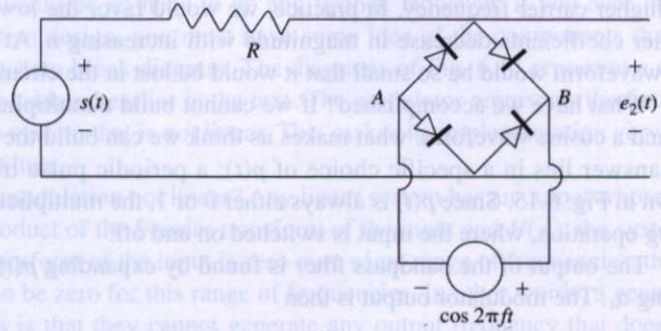
Equation (6.14) has been written for a gating function that spends half of its time high and half at zero. In fact, an AM wave will be produced for any value of the *duty cycle*.

The gating function can be implemented either passively or actively. Figure 6.16 shows two passive implementations. Figure 6.16(a) shows a simple switch that periodically shorts out the input. When the switch is open, the output equals the input. When the switch is closed, the output is zero. The resistance is the resistance of the source. The disadvantage of mechanical switching is that S must switch at a rate equal to the carrier frequency (or a submultiple of it if we select a harmonic). If our carrier frequency is in the megahertz range, mechanical switching is not practical.

Figure 6.16(b) presents a variation of the switch circuit in which the switching is accomplished using a *diode bridge* circuit. When $\cos 2\pi ft$ is positive, the point labeled B is at a higher potential than the point labeled A. In this condition, all four (ideal and matched) diodes are open circuited, and the circuit is equivalent to that of Fig. 6.16(a) with the switch open. On the other hand, when $\cos 2\pi ft$ is negative, point A is at a

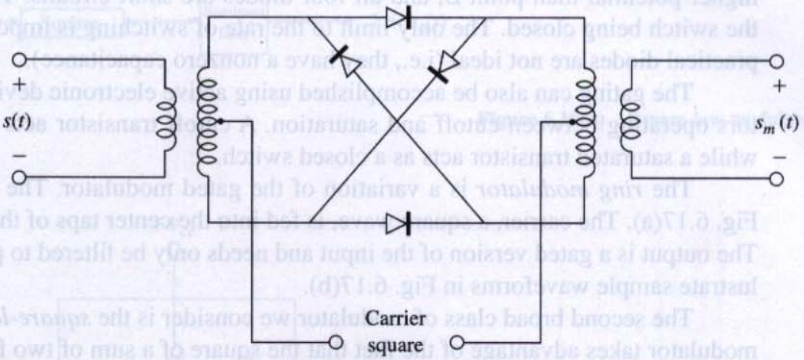


(a)

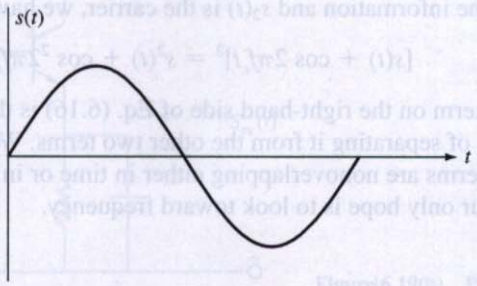


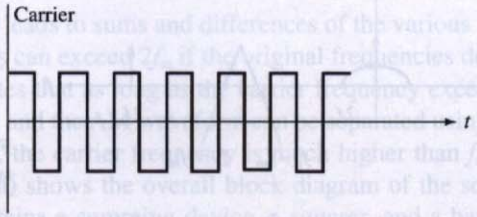
(b)

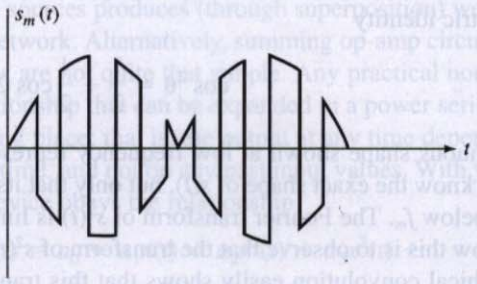
Figure 6.16 Implementation of gating function.



(a) Modulator







(b) Waveforms

Figure 6.17 Ring Modulator.

higher potential than point B, and all four diodes are short circuits. This is equivalent to the switch being closed. The only limit to the rate of switching is imposed by the fact that practical diodes are not ideal (i.e., they have a nonzero capacitance).

The gating can also be accomplished using active electronic devices such as transistors operating between cutoff and saturation. A cutoff transistor acts as an open switch, while a saturated transistor acts as a closed switch.

The *ring modulator* is a variation of the gated modulator. The circuit is shown in Fig. 6.17(a). The carrier, a square wave, is fed into the center taps of the two transformers. The output is a gated version of the input and needs only be filtered to produce AM. We illustrate sample waveforms in Fig. 6.17(b).

The second broad class of modulator we consider is the *square-law modulator*. This modulator takes advantage of the fact that the square of a sum of two functions contains a cross-product term that is the product of the two functions. That is,

$$[s_1(t) + s_2(t)]^2 = s_1^2(t) + s_2^2(t) + 2s_1(t)s_2(t) \tag{6.15}$$

If $s_1(t)$ is the information and $s_2(t)$ is the carrier, we have

$$[s(t) + \cos 2\pi f_c t]^2 = s^2(t) + \cos^2 2\pi f_c t + 2s(t)\cos 2\pi f_c t \tag{6.16}$$

The third term on the right-hand side of Eq. (6.16) is the desired AM waveform. We must find a way of separating it from the other two terms. We know that separation will be simple if the terms are nonoverlapping either in time or in frequency. Clearly, they overlap in time, so our only hope is to look toward frequency.

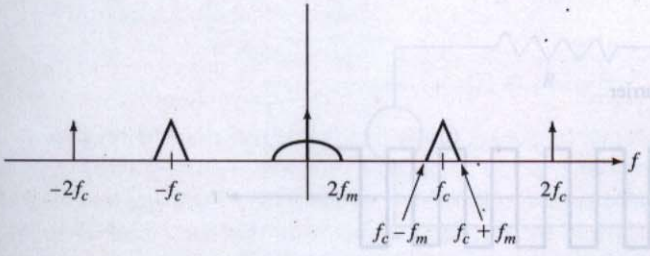


Figure 6.18 Fourier transform of squared-sum signal.

Figure 6.18 shows the Fourier transform of the signal in Eq. (6.16). The impulses at the origin and at $2f_c$ result from expanding the square of the cosine by means of the trigonometric identity

$$\cos^2\theta = \frac{1}{2} + \frac{1}{2}\cos 2\theta \tag{6.17}$$

The continuous shape shown at low frequency represents the Fourier transform of $s^2(t)$. We do not know the exact shape of $s(t)$, but only that its Fourier transform is limited to frequencies below f_m. The Fourier transform of $s^2(t)$ is limited to frequencies below $2f_m$. One way to show this is to observe that the transform of $s^2(t)$ is the convolution of $S(f)$ with itself. Graphical convolution easily shows that this transform goes to zero at $2f_m$. Another way to show it is to consider $s(t)$ as a sum of individual sinusoids at frequencies below f_m. When this sum is squared, the result is all possible cross products of terms. Trigonometric

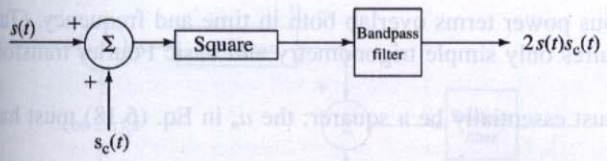


Figure 6.19(a) Square-law modulator.

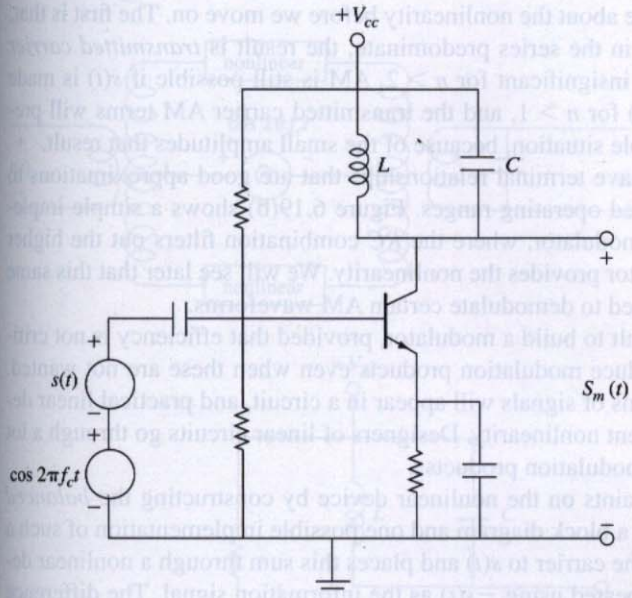


Figure 6.19(b) Practical square-law modulator.

identities tell us that this leads to sums and differences of the various frequencies. None of these sums or differences can exceed $2f_m$ if the original frequencies do not exceed f_m.

Figure 6.18 indicates that as long as the carrier frequency exceeds $3f_m$, the terms do not overlap in frequency, and the AM waveform can be separated using a bandpass filter. In most practical situations, the carrier frequency is much higher than f_m, so the condition is easily met. Figure 6.19(a) shows the overall block diagram of the square-law modulator. This block diagram contains a summing device, a squarer, and a bandpass filter. You already know how to build a bandpass filter. Summing devices can be active or passive. Any resistive circuit with two sources produces (through superposition) weighted sums of these sources throughout the network. Alternatively, summing op-amp circuits can be used.

Square-law devices are not quite that simple. Any practical nonlinear device has an output-versus-input relationship that can be expanded in a power series. This assumes that no energy storage is taking place; that is, the output at any time depends only on the value of the input at that same time, and not on any past input values. With $y(t)$ as output and $x(t)$ as input, the nonlinear device obeys the relationship

$$y(t) = a_0 + a_1x(t) + a_2x^2(t) + a_3x^3(t) + \cdots \tag{6.18}$$

The term we are interested in is $a_2x^2(t)$. If we could somehow find a way of separating this term from all the others, the nonlinear device could be used as a squarer.

Unfortunately, the various power terms overlap both in time and frequency. (Take the time to verify this! It requires only simple trigonometry and basic Fourier transform theory.)

The nonlinear device must essentially be a squarer; the a_n in Eq. (6.18) must have the property that

$$a_n << a_2, \quad \text{for } n > 2$$

There are several things to note about the nonlinearity before we move on. The first is that, if the $n = 1$ and $n = 2$ terms in the series predominate, the result is *transmitted carrier AM*. Further, if the a_n are *not* insignificant for $n > 2$, AM is still possible if $s(t)$ is made very small. Then $s^n(t) << s(t)$ for $n > 1$, and the transmitted carrier AM terms will predominate. This is not a desirable situation, because of the small amplitudes that result.

Semiconductor diodes have terminal relationships that are good approximations to square-law devices over limited operating ranges. Figure 6.19(b) shows a simple implementation of the square-law modulator, where the *RC* combination filters out the higher frequency term and the transistor provides the nonlinearity. We will see later that this same circuit configuration can be used to demodulate certain AM waveforms.

In reality, it is not difficult to build a modulator, provided that efficiency is not critical. In fact, most circuits produce modulation products even when these are not wanted. Superposition dictates that sums of signals will appear in a circuit, and practical linear devices always have some inherent nonlinearity. Designers of linear circuits go through a lot of effort to reduce unwanted modulation products.

We can relax the constraints on the nonlinear device by constructing the *balanced modulator*. Figure 6.20 shows a block diagram and one possible implementation of such a modulator. This system adds the carrier to $s(t)$ and places this sum through a nonlinear device. The operation is then repeated using $-s(t)$ as the information signal. The difference of the two outputs is taken, resulting in a cancellation of the terms due to odd powers in the expansion of Eq. (6.18). We illustrate the process by examining the cubic term in the equation. When we expand

$$[s(t) + \cos 2\pi f_c t]^3$$

the term that overlaps the frequency band of the AM waveform is

$$s^2(t)\cos 2\pi f_c t$$

This term remains unchanged when $-s(t)$ is substituted for $s(t)$. The result is that it cancels from the output of the balanced system. The desired term, $s(t)\cos 2\pi f_c t$, changes sign when $-s(t)$ is substituted for $s(t)$. Therefore, the operation of taking the difference has the effect of doubling the desired term. A similar approach is used to show that higher odd powers do not create undesired terms in the output. The balanced modulator is particularly effective if the nonlinearity has strong linear, square, and cubic terms, and all higher terms in the series are negligible. We should note that, since the first-order term is eliminated, the output of the balanced modulator is *suppressed carrier AM*.

A practical implementation of a square-law modulator is shown in Fig. 6.21. This common-emitter transistor circuit uses the nonlinearity of the transistor to produce the product of the signal and the carrier. The tuned circuit in the collector filters out the undesired harmonics.

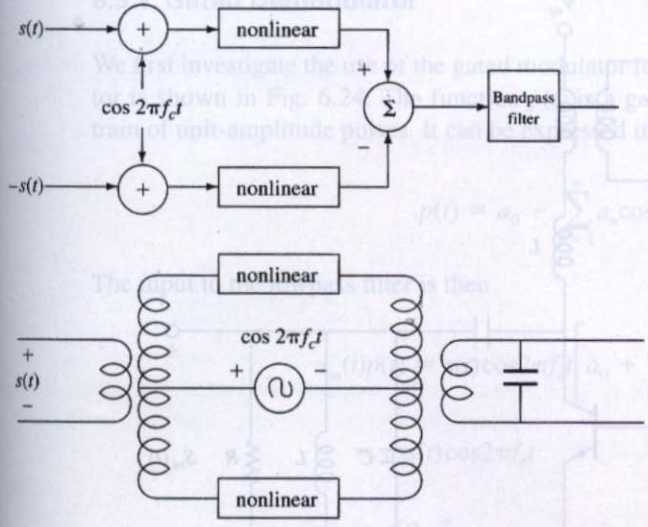


Figure 6.20 Balanced modulator.

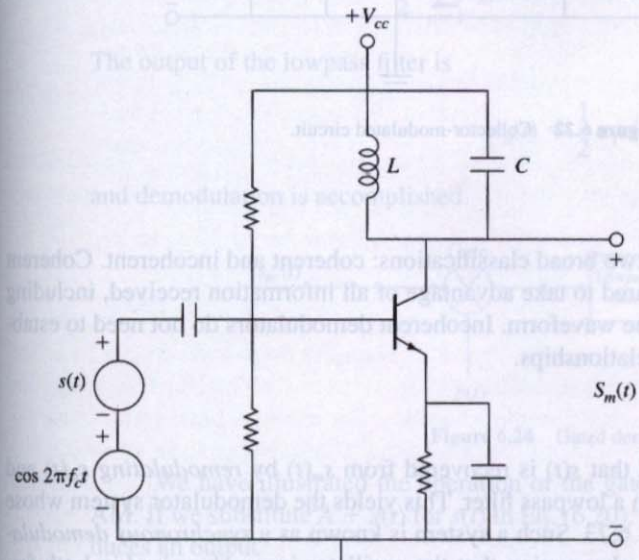


Figure 6.21 Implementation of square-law modulator.

The *waveshape modulator* can be thought of as a brute-force device. If you wanted to modulate a flow of water in a hose, you could hold your hand on the valve and keep turning it back and forth. An analogy to this simple system exists in electronics. You can envision building a power amplifier (or oscillator) that produces the carrier. Then simply vary the supply voltage to this amplifier in a manner that follows the information signal. The collector-modulated circuit of Fig. 6.22 does just this. The waveform appearing at the top of the *RLC* tuned collector circuit is the sum of V_{CC} and the information signal. We are therefore essentially varying the supply voltage in accordance with $s(t)$. The output is bandpass filtered to eliminate harmonics created by nonlinear operations of the transistor.

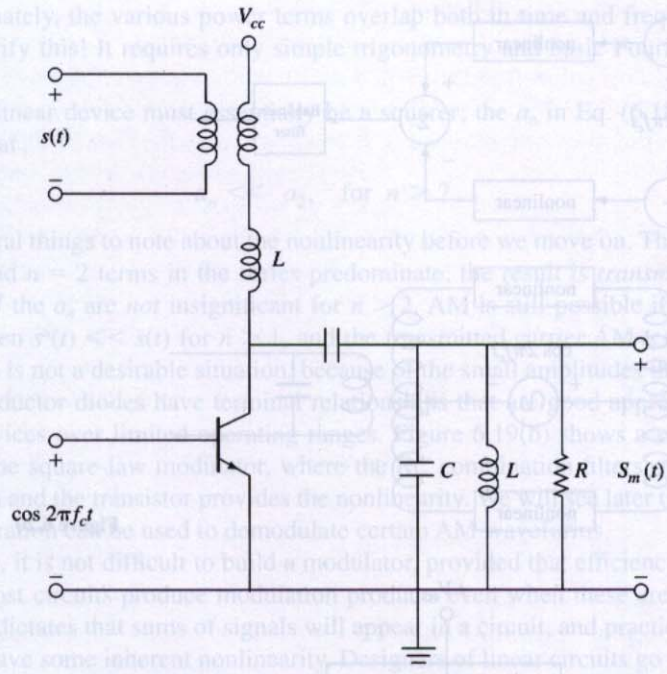


Figure 6.22 Collector-modulated circuit.

6.5 DEMODULATORS

We divide demodulators into two broad classifications: coherent and incoherent. Coherent demodulators must be configured to take advantage of all information received, including the amplitude and timing of the waveform. Incoherent demodulators do not need to establish absolute timing (phase) relationships.

Coherent Demodulation

We have previously observed that $s(t)$ is recovered from $s_m(t)$ by *remodulating* $s_m(t)$ and then passing the result through a lowpass filter. This yields the demodulator system whose block diagram appears in Fig. 6.23. Such a system is known as a *synchronous demodulator*. It gets its name from the observation that the oscillator is *synchronized* in both frequency and phase with the received carrier. Since the multiplier in the figure looks no different from the multiplier used in the modulator, we might expect variations of the gated and square-law modulators to be applicable.

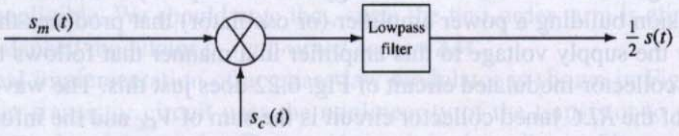


Figure 6.23 AM demodulator.

6.5.1 Gated Demodulator

We first investigate the use of the gated modulator for demodulation. The gated demodulator is shown in Fig. 6.24. The function $p(t)$ is a gating function consisting of a periodic train of unit-amplitude pulses. It can be expressed in a Fourier series as

$$p(t) = a_0 + \sum_{n=1}^{\infty} a_n \cos 2\pi n f_c t \tag{6.19}$$

The input to the lowpass filter is then

$$s_m(t)p(t) = s(t)\cos 2\pi f_c t \left[a_0 + \sum_{n=1}^{\infty} a_n \cos 2\pi n f_c t \right]$$

$$= a_0 s(t)\cos 2\pi f_c t \tag{6.20}$$

$$+ \frac{s(t)}{2} \sum_{n=1}^{\infty} a_n [\cos(n-1)2\pi f_c t + \cos(n+1)2\pi f_c t]$$

The output of the lowpass filter is

$$s_o(t) = \frac{1}{2} a_1 s(t) \tag{6.21}$$

and demodulation is accomplished.

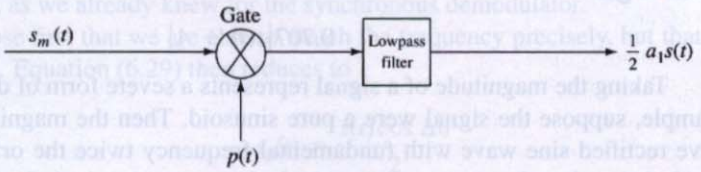


Figure 6.24 Gated demodulator.

We have illustrated the operation of the gated demodulator for suppressed carrier AM. If we substitute $A + s(t)$ for $s(t)$ in Eq. (6.20), we see that the gated demodulator produces an output

$$s_o(t) = \frac{1}{2} a_1 [A + s(t)] \tag{6.22}$$

This represents the original information signal, shifted by an amplitude constant. If the system contains ac-coupled devices, the constant will not appear in the output. If all amplifiers in the system are dc coupled, we may wish to remove the constant using a relatively large series capacitor that charges to the average value of the signal. We are assuming that the average value of the information, $s(t)$, is zero. If this were not true, removal of the constant would also remove some of the signal. Fortunately, most $s(t)$ information signals have zero dc value.

6.5.2 Square-Law Demodulator

We investigate the effect of adding the AM wave to a pure carrier term and then squaring the sum. This yields

$$[s_m(t) + A\cos 2\pi f_c t]^2 \tag{6.23}$$

Equation (6.23) can be rewritten as

$$\begin{aligned}\{[s(t) + A]\cos 2\pi f_c t\}^2 &= [s(t) + A]^2 \cos^2 2\pi f_c t \\ &= \frac{[s(t) + A]^2 + [s(t) + A]^2 \cos 4\pi f_c t}{2}\end{aligned} \tag{6.24}$$

The second term in Eq. (6.24) is an AM wave with a carrier frequency of $2f_c$ Hz. It can therefore be easily rejected by a lowpass filter. The first term can be expanded as

$$\frac{s^2(t)}{2} + \frac{A^2}{2} + As(t) \tag{6.25}$$

Unfortunately, the frequency content of $s^2(t)$ overlaps that of $s(t)$, and the two terms cannot be separated. However, suppose we used a lowpass filter to isolate the entire term

$$\frac{[s(t) + A]^2}{2} \tag{6.26}$$

from Eq. (6.24). Note that this lowpass filter must pass frequencies up to $2f_m$. We have then recovered the square of the sum of A and $s(t)$. We could subsequently take the square root of this to get

$$0.707|s(t) + A| \tag{6.27}$$

Taking the magnitude of a signal represents a severe form of distortion. As a simple example, suppose the signal were a pure sinusoid. Then the magnitude would be a full-wave rectified sine wave with fundamental frequency twice the original frequency. The rectified signal no longer contains a single frequency, but includes harmonics. If we listened to their sound in a speaker, the original sinusoid would be a pure tone, while the full-wave rectified sine wave would be a raspy tone one octave higher, due to the harmonic content. If the original signal were composed of a mixture of many frequencies, the distortion effect would be far more severe. Indeed, full-wave rectified voice is not intelligible. (Try it in the lab!)

But suppose A is large enough such that $s(t) + A$ never goes negative. In that case, the magnitude of $s(t) + A$ is equal to $s(t) + A$, and we have accomplished demodulation. This means that the added carrier at the receiver must have an amplitude greater than or equal to the maximum negative excursion of $s(t)$.

Effects of Frequency Mismatch

The demodulators we have been discussing require that we generate a replica of the carrier at the receiver. The replica must be synchronized with the received carrier. (Frequency and phase must be matched.) Let us investigate the consequences of frequency and phase mis-

matches. We illustrate the phenomenon for a suppressed carrier AM wave. Suppose that the local oscillator of Fig. 6.23 is mismatched in frequency by Δf and in phase by $\Delta\theta$. The output of the multiplier is then

$$\begin{aligned} & s_m(t)\cos[2\pi(f_c + \Delta f)t + \Delta\theta] \\ & \quad = s(t)\cos 2\pi f_c t \cos[2\pi(f_c + \Delta f)t + \Delta\theta] \\ & \quad = s(t)\left[\frac{\cos[2\pi\Delta f t + \Delta\theta]}{2} + \frac{\cos[2\pi(2f_c + \Delta f)t + \Delta\theta]}{2}\right] \end{aligned} \tag{6.28}$$

Since Eq. (6.28) forms the input to the lowpass filter of the synchronous demodulator, the output of this filter is

$$s_o(t) = s(t)\frac{\cos(2\pi\Delta f t + \Delta\theta)}{2} \tag{6.29}$$

This results because the second term of Eq. (6.28) has frequency content around $2f_c + \Delta f$ and is therefore rejected by the lowpass filter. Equation (6.29) represents a signal $s(t)$ multiplied by a sinusoid at Δf Hz. We can assume that Δf is small, since we attempt to make it equal to zero. The modulation theorem then indicates that $s_o(t)$ has a Fourier transform with frequencies ranging up to $f_m + \Delta f$. Thus, even though the lowpass filter is designed to pass frequencies only up to f_m, it is reasonable to assume that this entire term passes through the filter (since $\Delta f << f_m$).

Note that if the phase and frequency are perfectly adjusted, Eq. (6.29) reduces simply to $s(t)/2$, as we already knew for the synchronous demodulator.

Suppose first that we are able to match the frequency precisely, but that the phase is mismatched. Equation (6.29) then reduces to

$$s_o(t) = \frac{s(t)\cos\Delta\theta}{2} \tag{6.30}$$

This is an undistorted version of $s(t)$, so we would normally not be concerned. However, as the phase mismatch approaches 90°, the output goes to zero. If noise is added to the signal, the attenuation presented by the $\cos\Delta\theta$ term could become a significant negative factor. That is, as $\Delta\theta$ deviates from zero, the signal to noise ratio decreases.

One method of making the receiver insensitive to phase variations (i.e., making it robust) is to use the *quadrature receiver*, as shown in Fig. 6.25. We have indicated a phase shift of $\Delta\theta$ on both the sine and cosine multiplier signal. Equivalently, we could have indicated this phase shift on the input carrier.

Using trigonometric identities, we find the outputs of the two lowpass filters to be

$$\begin{aligned} s_1(t) &= \frac{1}{2}s(t)\cos\Delta\theta \\ s_2(t) &= -\frac{1}{2}s(t)\sin\Delta\theta \end{aligned} \tag{6.31}$$

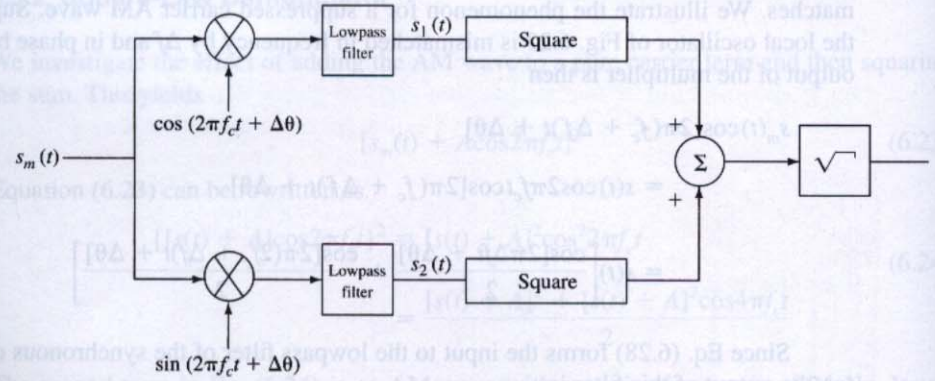


Figure 6.25 Quadrature receiver.

After taking the square root of the sum of the squares, we find that

$$s_o(t) = \frac{1}{2}\sqrt{s^2(t)} = \frac{1}{2}|s(t)| \tag{6.32}$$

As in the case of the square-law demodulator, undistorted demodulation is possible only if $s(t) \geq 0$. This means that the quadrature demodulator works only for transmitted carrier AM. Indeed, for that mode of transmission, we will find much simpler ways to demodulate a signal. We present the quadrature demodulator only to develop this important building block for later application.

Let us return to Eq. (6.29) and assume now that the phase has been perfectly matched, but that the frequency is mismatched. The output of the synchronous demodulator is then

$$s_o(t) = \frac{s(t)\cos 2\pi\Delta ft}{2} \tag{6.33}$$

The frequency mismatch is usually small (we try to make it zero), so the result will be a slowly varying amplitude (beating) of $s(t)$. If, for example, $s(t)$ is an audio signal and the frequency mismatch is 1 Hz, the effect would be to multiply $s(t)$ by a 1-Hz sinusoid. This is like taking the volume control of your radio and smoothly varying it from zero to maximum twice each second! Clearly, it is totally unacceptable. With a carrier frequency of 1 MHz, the 1-Hz mismatch represents only one part in 10^6. But suppose you were an expert at frequency matching, and your mismatch was only 10^{-3} Hz. Then your volume goes from maximum to zero once every 500 seconds. Unless we can derive the exact carrier from the incoming wave, or unless both the transmitter and receiver carriers are derived from the same source, synchronous demodulation is doomed to highly limited use.

6.5.3 Carrier Recovery in Transmitted Carrier AM

We have seen that synchronous demodulation requires *perfect* matching of the frequency and a phase mismatch that does not approach 90°. Frequency matching is possible if the AM waveform contains a periodic component at the carrier frequency. That is, the Fourier

transform of the received AM waveform must contain an impulse at the carrier frequency. This is the case with transmitted carrier AM.

We assume that the received signal is of the form

$$s(t)\cos 2\pi f_c t + A\cos 2\pi f_c t \tag{6.34}$$

One way to extract the carrier is with a very narrow bandpass filter tuned to the carrier frequency. In the steady state, all of the carrier term will pass through this filter, while only a portion of the modulated carrier will go through. The Fourier transform of the filter output is

$$S_o(t) = \frac{S(f - f_c) + S(f + f_c) + A\delta(f + f_c) + A\delta(f - f_c)}{2} \tag{6.35}$$

This equation applies for the range of frequencies in the passband of the filter, that is,

$$f_c - \text{BW}/2 < f < f_c + \text{BW}/2$$

The inverse transform is then

$$s_o(t) = A\cos 2\pi f_c t + \int_{f_c - \text{BW}/2}^{f_c + \text{BW}/2} S(f - f_c)\cos 2\pi f t df \tag{6.36}$$

The integral in Eq. (6.36) is bounded by

$$\frac{1}{2\pi t} S_{\max}(f)BW \tag{6.37}$$

The smaller the bandwidth of the filter, the closer is the output to the pure carrier term.

An alternative to the narrow filter is a *phase-lock loop,* illustrated in Fig. 6.26. The phase-lock loop is discussed in detail in the next chapter. For now, we merely indicate that, if properly designed, the loop will lock on to the periodic component in the input to produce a sinusoid at the carrier frequency.

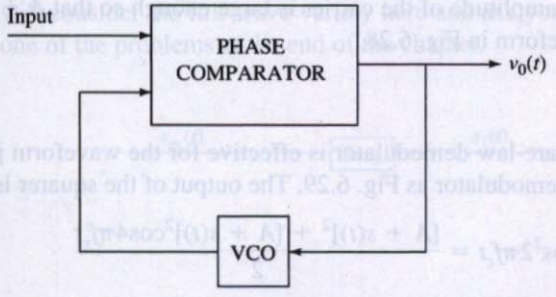


Figure 6.26 The phase-lock loop.

The implementation of the synchronous detector for transmitted carrier AM is shown in Fig. 6.27. Figure 6.27(a) shows the bandpass filter used for carrier recovery, and Fig. 6.27(b) shows the phase-lock loop.

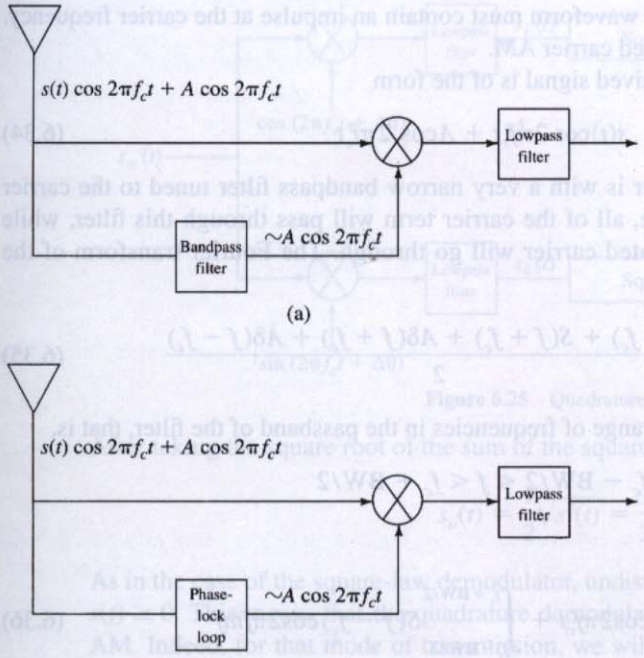


Figure 6.27 Carrier recovery in transmitted carrier AM.

6.5.4 Incoherent Demodulation

Coherent demodulators (detectors) require reproduction of the carrier at the receiver. Since the exact carrier frequency and phase must be matched at the detector, accurate timing information is needed.

If the carrier term is sufficiently large in transmitted carrier AM, it is possible to use incoherent detectors that do not have to reproduce the carrier or determine timing information. Let us suppose that the amplitude of the carrier is large enough so that $A + s(t) \geq 0$. We sketch a typical AM waveform in Fig. 6.28.

Square-law Detector

We noted earlier that the square-law demodulator is effective for the waveform presented in Fig. 6.28. We repeat that demodulator as Fig. 6.29. The output of the squarer is

$$[A + s(t)]^2\cos^2 2\pi f_c t = \frac{[A + s(t)]^2 + [A + s(t)]^2\cos 4\pi f_c t}{2} \tag{6.38}$$

The output of the lowpass filter (which passes frequencies up to $2f_m$) is

$$s_1(t) = \frac{[A + s(t)]^2}{2} \tag{6.39}$$

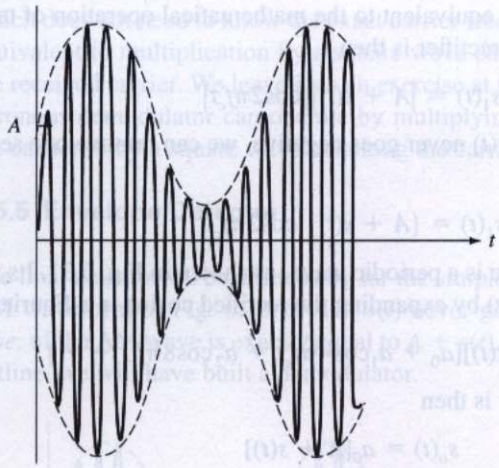


Figure 6.28 Transmitted carrier AM where $A + s(t) \geq 0$.

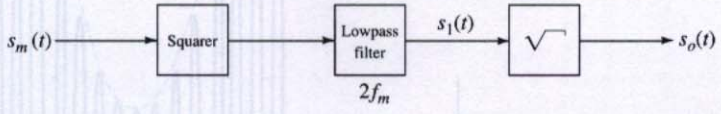


Figure 6.29 Square-law detector for transmitted carrier AM.

If we now assume that A is large enough such that $A + s(t)$ never goes negative, the output of the square rooter is

$$s_o(t) = 0.707[A + s(t)] \tag{6.40}$$

and demodulation is accomplished.

Rectifier detector

The squarer can be replaced by other forms of nonlinearity. In particular, consider the rectifier detector shown in Fig. 6.30. The rectifier can be either half wave or full wave. We will consider the full-wave variety here and ask you to examine the half-wave rectifier in one of the problems at the end of the chapter.

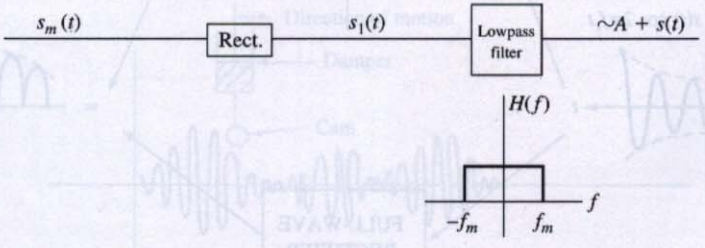


Figure 6.30 Rectifier detector.

Full-wave rectification is equivalent to the mathematical operation of taking the absolute value. The output of the rectifier is then

$$s_1(t) = |A + s(t)||\cos 2\pi f_c t| \tag{6.41}$$

but since we assume that $A + s(t)$ never goes negative, we can remove one set of absolute value signs to get

$$s_1(t) = \{A + s(t)\}|\cos 2\pi f_c t| \tag{6.42}$$

The absolute value of the cosine is a periodic wave, as shown in Fig. 6.31. Its fundamental frequency is $2f_c$. We rewrite $s_1(t)$ by expanding the rectified cosine in a Fourier series:

$$s_1(t) = [A + s(t)][a_0 + a_1 \cos 4\pi f_c t + a_2 \cos 8\pi f_c t + \cdots] \tag{6.43}$$

The output of the lowpass filter is then

$$s_o(t) = a_0[A + s(t)] \tag{6.44}$$

and demodulation is accomplished.

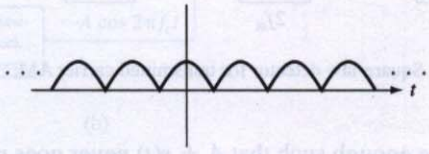


Figure 6.31 Rectified sine wave.

Before leaving the rectifier detector, we will point out the mechanism by which this detector reconstructs the carrier waveform. Figure 6.32 shows that full-wave rectification of the AM wave is equivalent to multiplying the waveform by a square wave at the carrier frequency. That is, the process of taking the absolute value flips around the negative portion of the carrier. This is equivalent to multiplication by -1. Therefore, the rectifier,

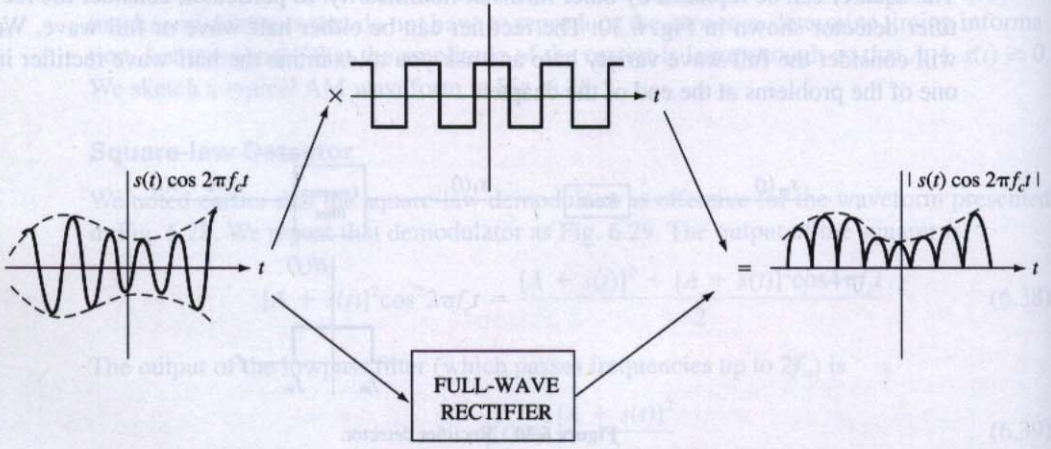


Figure 6.32 Multiplication by square wave vs. full-wave rectification.

which does not need to know the exact carrier frequency, is performing an operation that is equivalent to multiplication by a square wave carrier at the exact frequency and phase of the received carrier. We leave it as an exercise at the end of the chapter to show that a synchronous demodulator can operate by multiplying the wave either by a cosine matching the carrier or by a square wave matching the carrier.

6.5.5 Envelope Detector

The final detector we examine is by far the simplest. Let us observe the transmitted carrier AM waveform of Fig. 6.33. If $A + s(t)$ never goes negative, the upper outline, or *envelope,* of the AM wave is exactly equal to $A + s(t)$. If we can build a circuit that follows this outline, we will have built a demodulator.

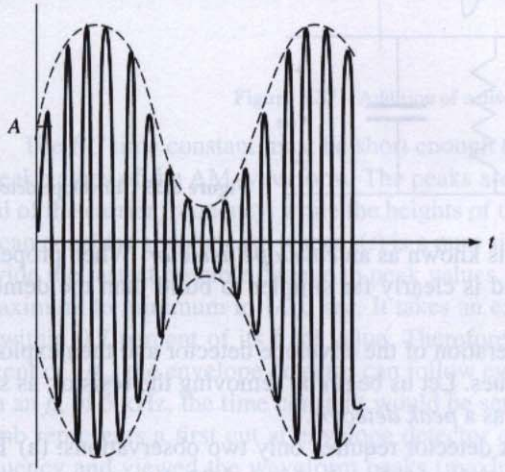


Figure 6.33 Transmitted carrier AM waveform.

It may be helpful to borrow an example from mechanics. Suppose that instead of representing a voltage waveform as a function of time, the curve represented the shape of a wire. One can envisage a cam moving along the top surface. If the cam is attached by means of a shock absorber, or viscous damper device, it will approximately follow the upper outline of the curve. This is shown in Fig. 6.34. The behavior is much the same as that of an automobile suspension system. You want the car to follow the outline of the road, but you do not wish it to track every ripple and bump in the road surface.

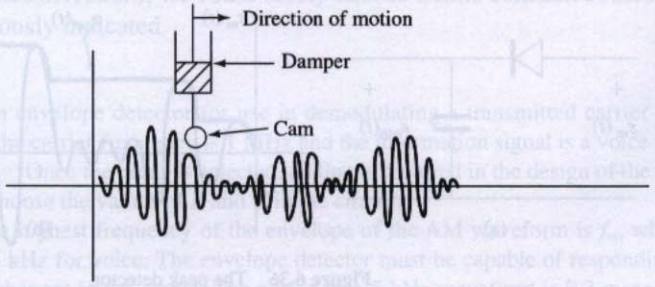


Figure 6.34 Mechanical outline follower.

The higher the carrier frequency, the more smoothly the cam will describe the upper outline of the curve, provided that it can respond fast enough to follow the shape of the outline. (You would probably not want your car to fly through the air between peaks of the road surface.) The outline, or envelope of the waveform, has a maximum frequency of f_m, while the ripples (the carrier) have a frequency of f_c. Intuition tells us that as long as $f_c >> f_m$, we can design the mechanical system.

We now construct the electrical analogy to this mechanical system. The mass of the cam is represented by a capacitor. ($F = ma = mv'$ is replaced by $i = Cv'$ for a capacitor.) The viscous friction provides a force proportional to velocity, much as a resistor provides a voltage proportional to current. Finally, the cam is not attached to the wire; the wire (road) can push upon the cam, but cannot pull it. This is a mechanical diode. The equivalent circuit is then as shown in Fig. 6.35.

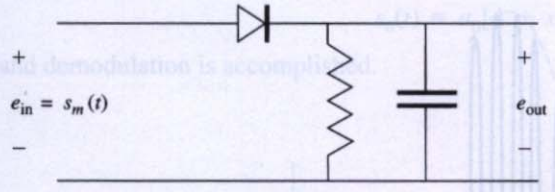


Figure 6.35 Envelope detector.

The circuit of Fig. 6.35 is known as an *envelope detector*. When properly designed, it serves as a demodulator and is clearly far simpler to build than the demodulators we have discussed previously.

We first examine the operation of the envelope detector and then explore the appropriate choice of parameter values. Let us begin by removing the resistor, as shown in Fig. 6.36(a). This circuit is known as a *peak detector*.

The analysis of the peak detector requires only two observations: (a) The input can never be greater than the output (for an ideal diode), and (b) the output can never decrease with time. The first observation is true because, if the input did exceed the output, the diode would be supporting a positive forward voltage. The second observation follows from the fact that the capacitor has no discharge path. Figure 6.36(b) shows a transmitted carrier AM waveform (the carrier frequency has been drawn much lower than it would be in practice, for illustrative purposes) and the output of the peak detector. The output is always equal to the maximum past value of the input.

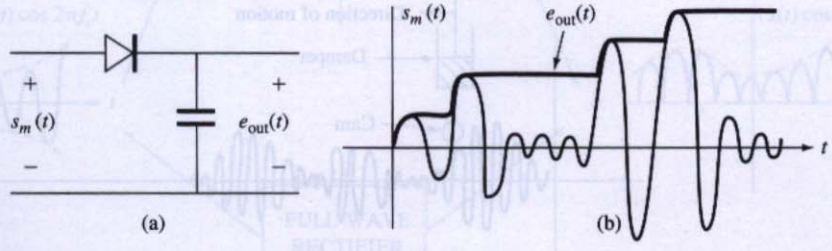


Figure 6.36 The peak detector.

If a discharging resistor is now added to the circuit, the output follows an exponential curve between peaks of the AM wave. This is shown in Fig. 6.37. If the time constant of the *RC* circuit is appropriately chosen, the output approximately follows the outline of the input curve, and the circuit acts as a demodulator. The output contains ripple at the carrier frequency (residual radio frequency), but this does not cause a problem, since we are interested only in frequencies below f_m.

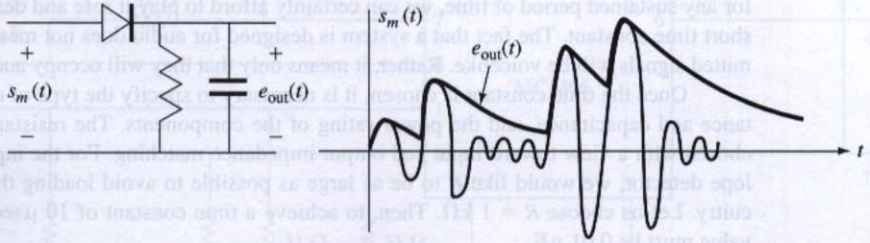


Figure 6.37 Addition of a discharging resistor.

The *RC* time constant must be short enough that the envelope can track the changes in peak values of the AM waveform. The peaks are spaced at intervals equal to the reciprocal of the carrier frequency, while the heights of these peaks follow the information, $s(t)$. We can consider a worst case where $s(t)$ is a pure sinusoid at a frequency of f_m. This would provide the fastest possible change in peak values. At this frequency, the peaks vary from a maximum to minimum in $1/2f_m$ sec. It takes an exponential function 5 time constants to get within 0.7 percent of its final value. Therefore, if we set the *RC* time constant to 10 percent of $1/f_m$, the envelope detector can follow even the highest frequency. For example, with an f_m of 5 kHz, the time constant would be set to 1/50 msec, or 20 μsec. This rule of thumb represents a first cut at envelope detector design. We chose the highest envelope frequency and viewed the waveform peaks (maximum and minimum). In reality, a sinusoid has its maximum slope in the middle and minimum slope at the extremes. Therefore, choosing the time constant on the basis of extremes means that our detector will not track all carrier peaks in between these extremes. We are saved by noting that typical information signals spend only a small fraction of their time at the highest frequency. Also, the large difference between carrier and envelope frequency allows a lot of leeway in choosing the time constant. If the information signal $s(t)$ is such that we expect significant periods near the highest frequency (as in the case, for example, of a soprano singing a particularly high-pitched selection), we could safely choose a time constant considerably smaller than that previously indicated.

Example 6.2

Design an envelope detector for use in demodulating a transmitted carrier AM waveform. Suppose the carrier frequency is 1 MHz and the information signal is a voice waveform.

Solution: Once the diode is selected, all that is required in the design of the envelope detector is to choose the values of *R* and *C* in the circuit.

The highest frequency of the envelope of the AM waveform is f_m, which we will assume is 5 kHz for voice. The envelope detector must be capable of responding to the fastest possible changes in the signal. The period of a 5-kHz waveform is 0.2 msec, and our guide-

line calls for choosing an RC time constant that is 10 percent of this value, or 20 μsec. This choice will not guarantee that the envelope detector output hits all of the peaks of the carrier. However, since a certain amount of ripple at 1 MHz will not hurt the system, we can afford to shorten the time constant. For example, a time constant of 10 μsec would allow the signal to come within 0.005 percent of the final value in tracking the fastest envelope frequency, and the carrier response would decrease only to 0.975 of the peak (i.e., the exponential decay over one period of the carrier). Even though the envelope is very rarely at the maximum frequency for any sustained period of time, we can certainly afford to play it safe and design for a rather short time constant. The fact that a system is designed for audio does not mean that all transmitted signals will be voicelike. Rather, it means only that they will occupy audio frequencies.

Once the time constant is chosen, it is necessary to specify the type of diode, the resistance and capacitance, and the power rating of the components. The resistance is normally chosen with a view toward input and output impedance matching. For the input to the envelope detector, we would like R to be as large as possible to avoid loading the previous circuitry. Let us choose $R = 1$ kΩ. Then, to achieve a time constant of 10 μsec, the capacitor value must be 0.01 μF.

6.5.6 Integrated Circuit Modulators and Demodulators

Several integrated circuit (IC) manufacturers have produced balanced modulators and demodulators. Among these are the Signetic MC1496/MC1596 and the Analog Devices AD630. These ICs contain differential amplifiers either that are driven into saturation or that simulate an electronic commutator (a device that alternately multiplies by positive and negative values). The reader may consult the literature for details of the electronics; we will concentrate upon applications in this text.

Figure 6.38 shows the MC1496 used as a transmitted carrier amplitude modulator. The same circuit can be used to generate suppressed carrier AM by choosing different resistor values in the carrier adjust circuitry.

The MC1496 is also used for demodulation of transmitted carrier AM. The circuit is shown in Fig. 6.39. The carrier for this operation is derived by driving the high-frequency amplifier into saturation, thereby providing an amplified and limited output that resembles a square wave at the carrier frequency. This carrier feeds into one of the MC1496 inputs, along with the AM wave into the other. The output must be lowpass filtered to recover the information signal.

6.6 BROADCAST AM AND SUPERHETERODYNE RECEIVER

The electromagnetic spectrum has been described as a natural resource. The dictionary defines *resource* as "something that lies ready to use or that can be drawn upon for aid or to take care of a need." *Natural resource* is defined as "those actual and potential forms of wealth supplied by nature." The electromagnetic spectrum does indeed lie ready for use to take care of a need—the need to communicate. Also, it is a potential form of wealth, and while it is not expendable in the sense in which oil and natural gas are, it does get used up in the sense that only a limited number of users can employ it at any one time.

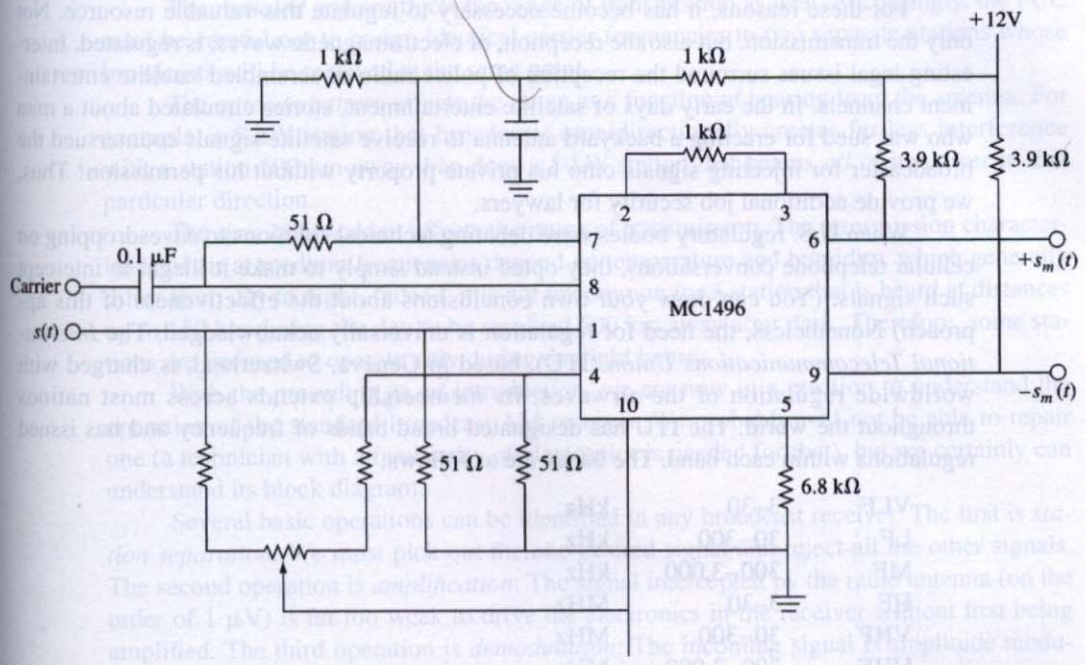


Figure 6.38 AM modulator.

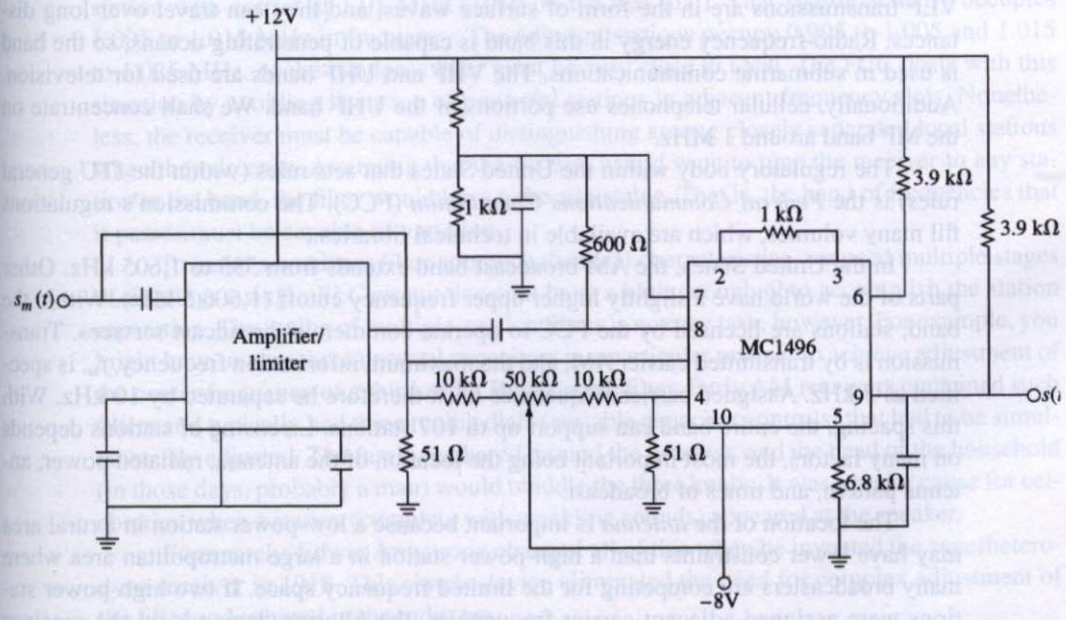


Figure 6.39 Transmitted carrier AM demodulator.

For these reasons, it has become necessary to regulate this valuable resource. Not only the transmission, but also the reception, of electromagnetic waves, is regulated. Interesting legal issues surround the reception of police radio or scrambled satellite entertainment channels. In the early days of satellite entertainment, stories circulated about a man who was sued for erecting a backyard antenna to receive satellite signals countersued the broadcaster for injecting signals onto his private property without his permission! Thus, we provide additional job security for lawyers.

When U.S. regulatory bodies were debating technical solutions to eavesdropping on cellular telephone conversations, they opted instead simply to make it illegal to intercept such signals. (You can draw your own conclusions about the effectiveness of this approach) Nonetheless, the need for regulation is universally acknowledged. The *International Telecommunications Union* (ITU), based in Geneva, Switzerland, is charged with worldwide regulation of the airwaves; its membership extends across most nations throughout the world. The ITU has designated broad bands of frequency and has issued regulations within each band. The bands are as follows:

VLF	3–30	kHz
LF	30–300	kHz
MF	300–3,000	kHz
HF	3–30	MHz
VHF	30–300	MHz
UHF	300–3,000	MHz
SHF	3–30	GHz
EHF	30–300	GHz

VLF transmissions are in the form of surface waves, and they can travel over long distances. Radio-frequency energy in this band is capable of penetrating oceans, so the band is used in submarine communications. The VHF and UHF bands are used for television. Additionally, cellular telephones use portions of the UHF band. We shall concentrate on the MF band around 1 MHz.

The regulatory body within the United States that sets rules (within the ITU general rules) is the *Federal Communications Commission* (FCC). The commission's regulations fill many volumes, which are available in technical libraries.

In the United States, the AM broadcast band extends from 535 to 1,605 kHz. Other parts of the world have a slightly higher upper frequency cutoff (1,606.5 kHz). Within the band, stations are licensed by the FCC to operate commercial broadcast services. Transmission is by transmitted carrier AM, and the maximum information frequency, f_m, is specified as 5 kHz. Assigned carrier frequencies must therefore be separated by 10 kHz. With this spacing, the entire band can support up to 107 stations. Licensing of stations depends on many factors, the most important being the location of the antenna, radiated power, antenna pattern, and times of broadcast.

The location of the *antenna* is important because a low-power station in a rural area may have fewer constraints than a high-power station in a large metropolitan area where many broadcasters are competing for the limited frequency space. If two high-power stations were assigned adjacent carrier frequencies, the filtering demands on the receiver would be excessive. The height and altitude of the antenna, of course, also affect the range of transmission.

The *radiated power* affects the range of transmission as well. Accordingly, the FCC must be careful not to assign identical carrier frequencies to two separate stations whose broadcasts will be received at the same point.

The *antenna pattern* affects the range as a function of bearing from the antenna. For example, a 5-kW station that broadcasts omnidirectionally creates far less interference with a station 500 km away than does a 5-kW station that beams *all* of its power in that particular direction.

The *time of broadcast* affects the range of transmission. The transmission characteristics of air at medium frequencies depend on temperature and humidity, which generally differ from day to night. Indeed, it is not uncommon for a station that is heard at distances up to 150 km during the day to be received 500 km away after dark. Therefore, some stations are licensed to operate only during daylight hours.

With the preceding as an introduction, we are now in a position to understand the operation of the standard broadcast AM receiver. We probably will not be able to repair one (a technician with a familiarity of electronics is needed for that), but we certainly can understand its block diagram.

Several basic operations can be identified in any broadcast receiver. The first is *station separation*: We must pick out the one desired signal and reject all the other signals. The second operation is *amplification*: The signal intercepted by the radio antenna (on the order of 1 μV) is far too weak to drive the electronics in the receiver without first being amplified. The third operation is *demodulation*: The incoming signal is amplitude modulated and contains frequencies centered about the carrier frequency.

Separating one channel from the others requires a very accurate bandpass filter with a sharp frequency cutoff characteristic. Suppose, for example, that we wish to listen to a station with carrier at 1.01 MHz (1010 on the AM dial). That station actually occupies 1.005 to 1.015 MHz in frequency. The adjacent stations occupy 0.995 to 1.005 and 1.015 to 1.025 MHz, so the bandpass filter must be very close to ideal. The FCC deals with this situation by avoiding licensing of powerful stations in adjacent frequency slots. Nonetheless, the receiver must be capable of distinguishing among closely separated local stations when they do exist. Assuming that the listener would want to tune the receiver to any station in the band, the filter would have to be adjustable. That is, the band of frequencies that it passes must be capable of variation.

To make a bandpass filter approach the ideal characteristics, we need multiple stages of filtering; a single *RLC* circuit does not have a high enough Q to accomplish the station separation. The tuning of multiple-section filters is no easy task, however. For example, you might have to vary three unequal capacitors in a particular manner to achieve adjustment of the center frequency of a third-order Butterworth filter. Early AM receivers contained such filters and typically had three tuning dials (variable capacitor controls) that had to be simultaneously adjusted. The family gathered around the *wireless*, and the head of the household (in those days, probably a man) would twiddle the three knobs. It was a major cause for celebration when a station (complete with crackling sounds) appeared at the speaker.

Fortunately, Edwin Armstrong changed all of this when he invented the superheterodyne receiver in 1918. This simple device eliminated the need for complex adjustment of the filter and ushered in the radio era.

Recall that multiplication of a signal by a sinusoid shifts all frequencies up and down by the frequency of the sinusoid. Because of this, station selection can be accom-

plished by building a *fixed* bandpass filter and shifting the input frequencies so that the station of interest falls in the passband of the filter. Looked at another way, we construct a viewing window on the frequency axis. Then, instead of moving this window around to view a particular portion of the axis, we keep the window stationary and shift the entire axis. The shifting process is known as *heterodyning*, and the resulting receiver is the *superheterodyne* receiver. The process of heterodyning is applicable to other forms of modulation (e.g., FM).

A block diagram of an AM broadcast receiver is shown in Fig. 6.40.

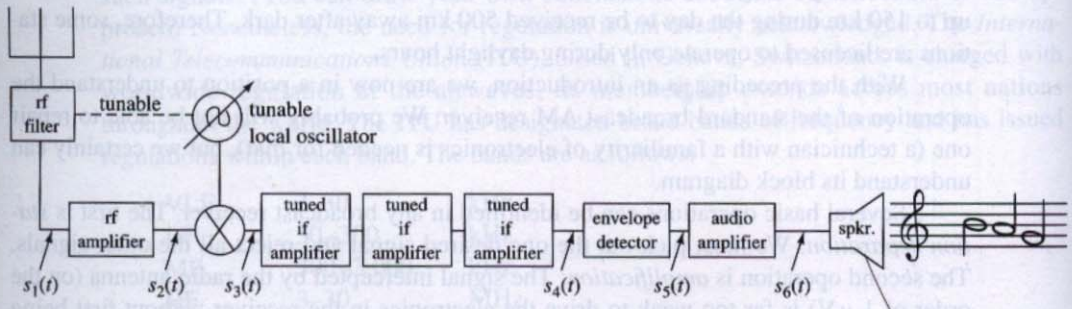


Figure 6.40 AM broadcast receiver.

In the figure, the antenna receives a signal that is a weighted sum of all broadcasted signals. After some filtering, which we will examine in a moment, the incoming signal is amplified in a radio frequency (rf) amplifier. The signal, $s_2(t)$, is then shifted up and down in frequency by multiplying by a sinusoidal generator, called the *local oscillator*. The shifting or heterodyning operation is also known as *mixing*.

The output of the heterodyner is applied to the sharp bandpass filter consisting of multiple filtering stages. This filter is normally combined with amplifiers. The fixed bandpass filter is set to 455 kHz, called the intermediate frequency (if), and has a bandwidth of 10 kHz, matching that of each station. This frequency is not within the AM broadcast band and is specified by the FCC. If stations were authorized to broadcast at this carrier, some of the signal would enter the if portion of the receiver (since every piece of wire acts as an antenna) and would be heard on top of the desired station. In most receivers, the if filter is made up of three tuned circuits that are aligned so as to generate a Butterworth filter characteristic (poles around a semicircle in the s-plane). At $s_4(t)$, we have a modulated signal whose carrier frequency has been shifted to 455 kHz and that has already been amplified and separated out from the other signals.

Let us now determine the required frequency of the local oscillator. Suppose you wish to listen to a station at the lower end of the dial, say, a carrier frequency of 540 kHz. To shift this frequency to 455 kHz, you would have to multiply by a sinusoid of either 85 or 995 kHz. Now suppose you wish to listen to a station at the top of the dial, with carrier at 1,600 kHz. Then the local oscillator setting must be either 1,145 or 2,055 kHz. To tune in any station in the band, the oscillator must be tunable over the range from 85 to 1,145 kHz or the range from 995 to 2,055 kHz.

The second of these ranges is selected for practical reasons. The local oscillator is set at the sum of 455 kHz and the desired carrier frequency. The resulting oscillator must tune over a range where the highest frequency is a little more than two times the lowest. If the first range had been selected, the highest frequency would be 13.5 times the lowest. It is much easier to construct variable oscillators for ranges that vary by a factor of 2 to 1 than by a factor of 13.5 to 1. The higher range might require a *range switch*.

The receiver then puts $s_4(t)$ through an envelope detector and then amplifies the signal (usually using push-pull power amplifiers) before applying it to a loudspeaker. After detection, the signal is at audio frequency (af). Some filtering may be done to provide treble and base controls.

One significant problem not mentioned earlier is that heterodyning produces both an upshift and a downshift in frequency (i.e., sums and differences). While one of these shifts moves the desired station into the if window (450 to 460 kHz), the other moves another band of frequencies into the same window. This undesired signal is called an *image*, and eliminating it is not very difficult.

As an example, suppose you wish to listen to a station at a carrier frequency of 600 kHz. Then the local oscillator is set to 600 + 455 = 1,055 kHz. Multiplication by this sinusoid places the desired 600-kHz station right into the if filter passband. But there is another station at a carrier of 1,055 + 455 = 1,510 kHz that, when multiplied by the local oscillator, will produce a component at 455 kHz. This *image station* would be heard right on top of the desired station.

The separation between the image and the desired station is twice the if frequency, or 910 kHz. This places the image 91 frequency slots away from the desired station. The 89 stations between the two are eliminated by the if filter.

A bandpass filter with a bandwidth of less than 1,820 kHz would accomplish the separation. Such a filter must pass the desired station while rejecting the station 910 kHz away. The filter must be tunable, but it need not be a sharp bandpass filter. We do not care what it does to the 89 stations between the image and the desired signal. A single tuned stage is therefore sufficient. This is shown in Fig. 6.41.

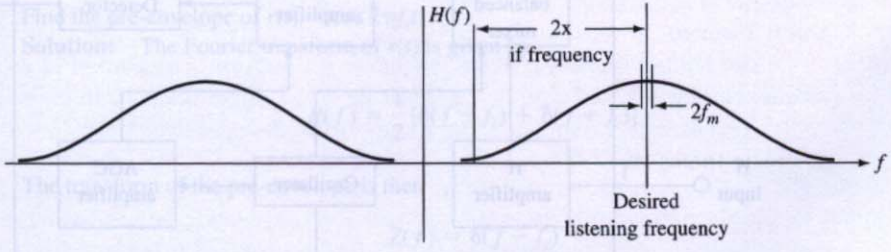


Figure 6.41 Image rejection process.

In practice, when the tuning dial on a receiver is turned, this sloppy RF rejection filter is tuned at the same time that the frequency of the local oscillator is changed. Before electronic tuning, the shaft of the tuning dial connected to two separate variable-capacitor

sections. One of these formed part of the image rejection filter, and the other formed part of the tuning circuit of the local oscillator. This is indicated in Fig. 6.40 as a dashed line connecting the two functions.

Integrated Circuit Receiver

Many of the functions of the superheterodyne AM radio receiver have been implemented using integrated circuits. One example is the TEA5550 AM radio circuit from Signetics. This chip contains the balanced mixer, if amplifier, detector, rf amplifier, local oscillator, and automatic gain control seen in the superheterodyne receiver. Its block diagram and pin configuration are shown in Fig. 6.42. We have discussed all of these functions except *automatic gain control* (AGC). As you tune a receiver across the AM dial, the various stations that come in have different voltage levels. The level depends on the transmitted power, the distance from the transmitter, the antenna pattern of both the transmitter and the receiver, and the frequency characteristics of the channel. To minimize the amount of adjustment of the volume control as we change from one station to another, and to decrease time-varying volume effects, a form of feedback is employed in most radio receivers. The output amplitude is sensed (using a lowpass filter with a time constant on the order of several seconds), and the level is fed back to the various amplifier stages to affect the Q points. As the level increase, the Q points are adjusted to reduce amplification, thus providing a form of negative feedback that tends to keep the output level constant.

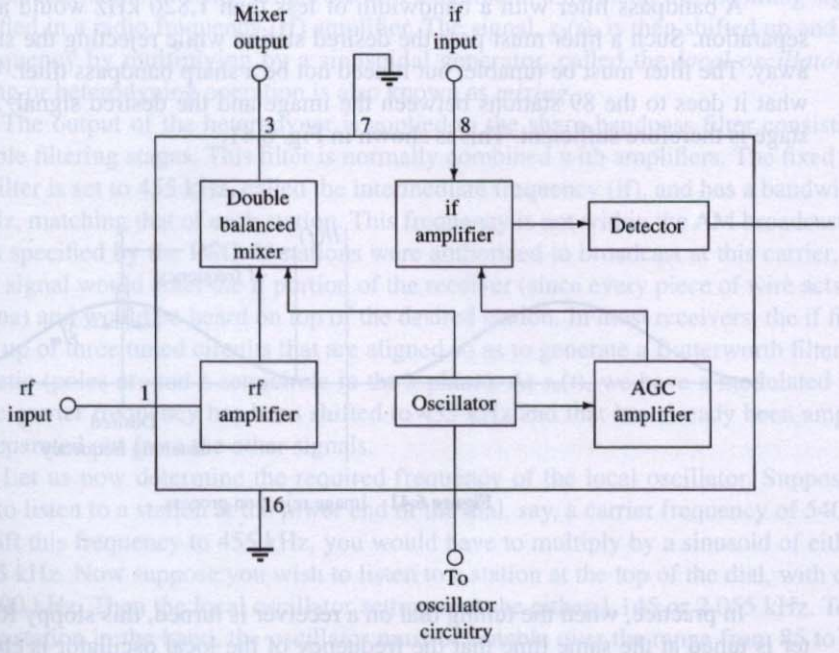


Figure 6.42 Integrated circuit AM receiver.

When the AM radio receiver chip is used as part of a radio, external circuitry is required. In particular, fabrication of capacitors of the size needed is not practical on the IC. We must therefore configure external circuitry for the oscillator and tuning portions of the system. (See application manuals from IC manufacturers for details of the actual circuitry for using the IC in a radio receiver.)

6.7 ENVELOPES AND PRE-ENVELOPES

We have referred to the *envelope* of a waveform as an outline that follows the peaks of the carrier. The envelope detector attempts to follow this outline. Our intuitive definition could be formalized as follows:

> Given $s(t)\cos 2\pi f_c t$, where the frequencies $s(t)$ are much less than f_c, the *envelope* is defined as the absolute value of $s(t)$.

This definition is based on observation of the time waveform.

The foregoing intuitive definition is not very satisfying, since it applies only to a narrow class of signals. We shall now explore a mathematical definition of the envelope that is identical to the intuitive definition for the class of functions to which the intuitive approach applies.

The *pre-envelope* (also known as the *complex envelope* or *analytic function*) of a waveform $r(t)$ is defined as the complex function of time whose Fourier transform is given by

$$Z(f) = 2U(f)R(f) \tag{6.45}$$

where $U(f)$ is the unit step function and $R(f)$ is the transform of $r(t)$. That is, the transform of the pre-envelope is zero for negative f and twice the original transform for positive f. The symbol $z(t)$ is commonly used to represent the pre-envelope function. Note that $z(t)$ cannot be a real function of time, since the magnitude of its transform is never even.

Example 6.3

Find the pre-envelope of $r(t) = \cos 2\pi f_c t$.

Solution: The Fourier transform of $r(t)$ is given by

$$R(f) = \frac{1}{2}[\delta(f - f_c) + \delta(f + f_c)]$$

The transform of the pre-envelope is then

$$Z(f) = \delta(f - f_c)$$

The function of time corresponding to this transform is

$$z(t) = e^{j2\pi f_c t}$$

You have seen this function in sinusoidal steady-state analysis in your circuits course. When you wish to find the response of a circuit to a sinusoid, you replace the sinusoid with a complex exponential. You do this primarily to carry the amplitude and phase in a single operation,

rather than having to track sines and cosines through the system. However, in reality, you are substituting the pre-envelope of the sinusoid for the actual function of time.

We now express $z(t)$ in terms of $r(t)$. That is, we translate the Fourier transform truncation process of Eq. (6.45) into an equivalent operation in the time domain. We start with

$$z(t) = 2r(t)*\mathcal{F}^{-1}[U(f)] \tag{6.46}$$

The inverse Fourier transform of $U(f)$ is found from the table in appendix II:

$$U(f) \leftrightarrow \frac{1}{2}\delta(f) - \frac{1}{2\pi jt} \tag{6.47}$$

Finally,

$$\begin{aligned} z(t) &= 2r(t)*\frac{1}{2}\delta(t) + 2r(t)*\frac{-1}{2\pi ft} \\ &= r(t) + \frac{j}{\pi}\int_{-\infty}^{\infty}\frac{r(\tau)}{t-\tau}d\tau \end{aligned} \tag{6.48}$$

Note that the real part of $z(t)$ is the original function of time, $r(t)$, just as the real part of $e^{j2\pi ft}$ is $\cos 2\pi ft$. The imaginary part of $z(t)$ is given by the convolution of $r(t)$ with $1/\pi t$. This convolution is useful in a number of applications and is known as the *Hilbert transform* of $r(t)$. The symbol for the Hilbert transform of a function of time is the same letter as that denoting the function, with the addition of a caret. Thus, the Hilbert transform of $r(t)$ is $\hat{r}(t)$.

The Fourier transform of $1/\pi t$ is $\text{sgn}(f)$. Therefore, taking the Hilbert transform of a function of time is equivalent to taking the mirror image of the portion of the Fourier transform to the left of the origin. Using this fact, we can see clearly how the pre-envelope results from Eq. (6.48). Taking the Fourier transform of that equation yields

$$Z(f) = R(f) + R(f)\text{sgn}(f) = R(f) + \hat{R}(f) \tag{6.49}$$

where $\hat{R}(f)$ is the Fourier transform of $\hat{r}(t)$. For positive f, $Z(f)$ is $2R(f)$, while for negative f, it is zero.

The Hilbert transform arises whenever we perform a truncation of a portion of the frequency axis. We will see it again when we study single sideband in the next section.

The *envelope* is defined as the magnitude of the pre-envelope.

Example 6.4

Find the envelope of $r(t) = \cos 2\pi f_c t$.

Solution: The pre-envelope of $r(t)$ was found in Example 6.3 to be

$$z(t) = e^{j2\pi f_c t}$$

The magnitude of this function is equal to unity. Therefore, the envelope of the function is a constant, 1. We already knew this from our intuitive definition of the envelope, since the function of time represents an unmodulated carrier wave.

Example 6.5

Find the envelope of

$$s_m(t) = s(t)\cos 2\pi f_c t$$

Solution: We must first find the pre-envelope of the function. The Hilbert transform is found by inverting the negative-f portion of the Fourier transform, $S_m(f)$:

$$\hat{S}_m(f) = S_m(f)\text{sgn}(f) = \frac{1}{2}[S(f - f_c) - S(f + f_c)]$$

In writing this equation, we have assumed that $S(f + f_c)$ lies completely to the left of the origin. This result therefore applies as long as $f_m < f_c$. The inverse Fourier transform of $S_m(f)$ is given by

$$\hat{s}_m(t) = js(t)\sin 2\pi f_c t$$

The pre-envelope is given by

$$z_m(t) = s(t)\cos 2\pi f_c t + js(t)\sin 2\pi f_c t$$

The envelope is the magnitude of the pre-envelope, or

$$|z_m(t)| = \sqrt{s^2(t)[\cos^2 2\pi f_c t + \sin^2 2\pi f_c t]}$$

$$= \sqrt{s^2(t)} = |s(t)|$$

This agrees with our intuitive definition. However, note that the intuitive definition requires that $f_c >> f_m$, while the mathematical definition requires only that $f_c > f_m$.

We have established a definition of envelope that applies to any function of time. A worthwhile, although very difficult, task would be to analyze the envelope detector circuit with general-input functions of time and compare its output to the (mathematical) envelope of the input.

The intuitive definition that applies when the envelope is slowly varying will prove sufficient for our work in AM. The concept of the pre-envelope proves useful in detection theory, which we explore in later chapters of the book dealing with digital communication.

6.8 SINGLE SIDEBAND (SSB)

In the AM systems we have studied, the range of frequencies required to transmit the signal is the band between $f_c - f_m$ and $f_c + f_m$. The frequency f_c is the carrier frequency, and f_m is the maximum frequency of the baseband signal $s(t)$. The total *bandwidth* is then $2f_m$. The frequency spectrum is a "natural resource" whose conservation is critical. The more frequency bandwidth required for each channel, the fewer number of stations can communicate simultaneously. Wouldn't it be lovely if we could find a way to send information using less than $2f_m$ of bandwidth?

Single sideband is a technique that allows transmission in half of the bandwidth required for AM double sideband. In Fig. 6.43, we define that portion of $S_m(f)$ which lies in the band above the carrier as the *upper sideband*. The portion below the carrier is the *lower sideband*. A double-sideband AM wave is composed of a lower and an upper sideband.

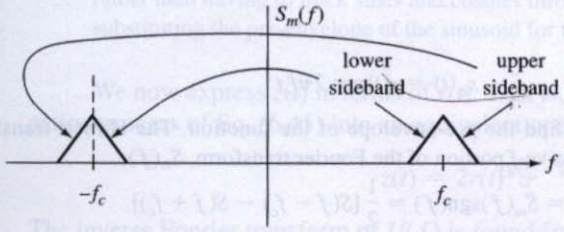


Figure 6.43 Definition of sidebands.

We can use the properties of the Fourier transform to show that the two sidebands are dependent on each other. The transform of the AM wave is formed by shifting $S(f)$, the transform of the signal, up and down in frequency. The lower sideband is formed from the negative-f portion of $S(f)$, and the upper sideband is the positive-f portion of $S(f)$. We assume that the information signal $s(t)$ is a real function of time. Therefore, the magnitude of $S(f)$ is even and the phase is odd. The negative-f portion of $S(f)$ can be derived from the positive-f portion by taking the complex conjugate. Similarly, the lower sideband of $s_m(t)$ can be derived from the upper sideband. Since the sidebands are not independent, it should be possible to transmit all essential information by sending only a single sideband. This is the essence of single-sideband communication.

Figure 6.44 shows the Fourier transforms of the upper and lower sideband versions of the AM wave, denoted $s_{\text{usb}}(t)$ and $s_{\text{lsb}}(t)$, respectively. The double-sideband AM wave is the sum of the two sidebands:

$$s_m(t) = s_{\text{lsb}}(t) + s_{\text{usb}}(t) \tag{6.50}$$

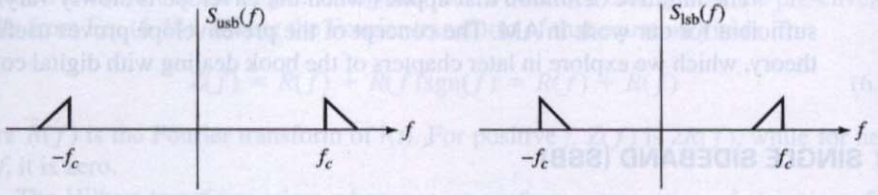


Figure 6.44 Single-sideband Fourier transforms.

Since the single-sideband waveform resides in a subset of the band of frequencies occupied by the double-sideband waveform, it automatically satisfies two of the requirements of a modulation system. That is, by proper choice of the carrier frequency, we can move the modulated waveform into a range of frequencies that transmits efficiently. We can also use different bands for different signals, thereby allowing simultaneous transmission of multiple signals.

The synchronous demodulator can be used to demodulate single sideband. This can be shown either pictorially in the frequency domain or mathematically in the time domain.

Looking first at frequencies, we know that multiplication by a sinusoid shifts the Fourier transform both up and down by the frequency of the sinusoid. Figure 6.45(a) shows the Fourier transform that results when $s_{usb}(t)$ is multiplied by a sinusoid at a frequency of f_c, and Fig. 6.45(b) shows a similar result for the lower sideband signal. In both cases, a low-pass filter would recover a replica of the original information signal.

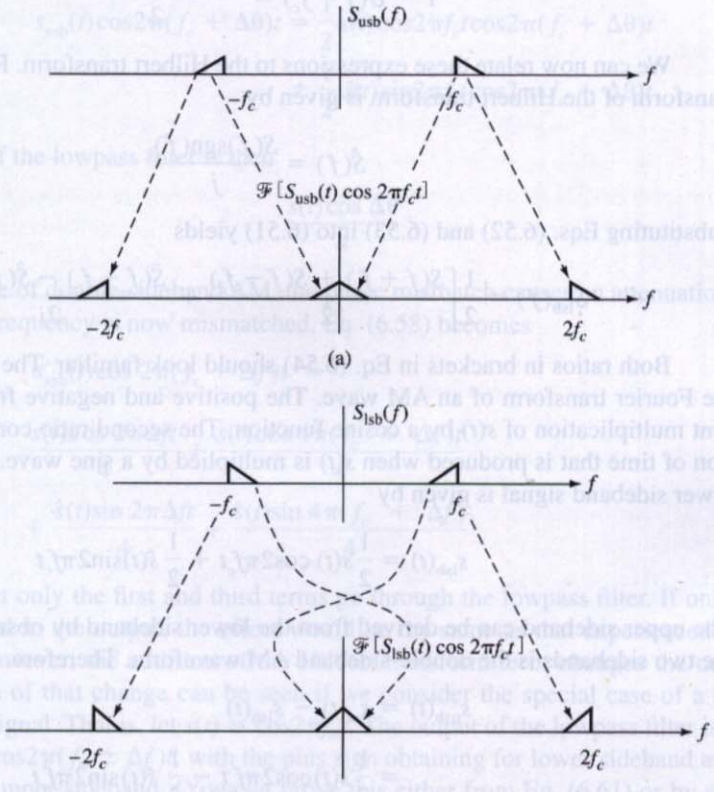


Figure 6.45 Demodulation of single sideband.

We can illustrate this same result in the time domain by multiplying the single-sideband waveform by a cosine at the same frequency as the carrier. Before continuing, however, we need to derive an expression for the single-sideband waveform that is a function of time.

We begin by expressing the lower sideband Fourier transform as a product of the double-sideband transform with a bandpass filter function. The filter passes only the lower sideband:

$$\begin{aligned} S_{lsb}(f) &= S_m(f)H(f) \\ &= \frac{S(f+f_c)U(f+f_c) - S(f-f_c)U(f-f_c) + S(f-f_c)}{2} \end{aligned} \tag{6.51}$$

We next express each unit step function of Eq. (6.51) in an alternate manner, using the *sign* (sgn) function:

$$U(f+f_c) = \frac{1+\text{sgn}(f+f_c)}{2} \qquad (6.52)$$
$$1-U(f-f_c) = \frac{1-\text{sgn}(f-f_c)}{2}$$

We can now relate these expressions to the Hilbert transform. Recall that the Fourier transform of the Hilbert transform is given by

$$\hat{S}(f) = \frac{S(f)\text{sgn}(f)}{j} \qquad (6.53)$$

Substituting Eqs. (6.52) and (6.53) into (6.51) yields

$$S_{\text{lsb}}(f) = \frac{1}{2}\left[\frac{S(f+f_c)+S(f-f_c)}{2} + \frac{\hat{S}(f-f_c)-\hat{S}(f+f_c)}{2j}\right] \qquad (6.54)$$

Both ratios in brackets in Eq. (6.54) should look familiar. The first is in the form of the Fourier transform of an AM wave. The positive and negative frequency shifts represent multiplication of $s(t)$ by a cosine function. The second ratio corresponds to the function of time that is produced when $s(t)$ is multiplied by a sine wave. The result is that the lower sideband signal is given by

$$s_{\text{lsb}}(t) = \frac{1}{2}s(t)\cos 2\pi f_c t + \frac{1}{2}\hat{s}(t)\sin 2\pi f_c t \qquad (6.55)$$

The upper sideband can be derived from the lower sideband by observing that the sum of the two sidebands is the double-sideband AM waveform. Therefore,

$$s_{\text{usb}}(t) = s_m(t) - s_{\text{lsb}}(t) \qquad (6.56)$$
$$= \frac{1}{2}s(t)\cos 2\pi f_c t - \frac{1}{2}\hat{s}(t)\sin 2\pi f_c t$$

Now that we have functions of time for the single-sideband waveforms, we can return to the analysis of the synchronous demodulator. We use the time domain expression of Eqs. (6.55) and (6.56) for the single-sideband waveforms:

$$s_{\text{ssb}}(t)\cos 2\pi f_c t = \frac{s(t)\cos^2 2\pi f_c t \pm \hat{s}(t)\sin 2\pi f_c t\cos 2\pi f_c t}{2} \qquad (6.57)$$

The plus sign applies to the lower sideband and the minus sign to the upper sideband. We use trigonometric expansions to express Eq. (6.57) as

$$s_{\text{ssb}}(t)\cos 2\pi f_c t = \frac{s(t) + s(t)\cos 4\pi f_c t \pm \hat{s}(t)\sin 4\pi f_c t}{4} \qquad (6.58)$$

The output of the lowpass filter with this quantity as input is simply $s(t)/4$, so we have accomplished demodulation.

Recall that in double sideband, we did not particularly favor the synchronous demodulator, since a phase mismatch led to attenuation and a frequency mismatch resulted in a very serious form of multiplicative distortion. Such a mismatch in single sideband is also serious, but slightly more forgiving than in the case of double sideband.

If the phase is mismatched, Eq. (6.58) becomes

$$\begin{aligned} s_{ssb}(t)\cos 2\pi(f_c + \Delta\theta)t = &\frac{1}{2}s(t)\cos 2\pi f_c t \cos 2\pi(f_c + \Delta\theta)t \\ &\pm \frac{1}{2}\hat{s}(t)\sin 2\pi f_c t \cos 2\pi(f_c + \Delta\theta)t \end{aligned} \tag{6.59}$$

The output of the lowpass filter is then

$$\frac{s(t)\cos \Delta\theta}{4} \tag{6.60}$$

As in the case of double-sideband AM, the phase mismatch causes an attenuation.

If the frequency is now mismatched, Eq. (6.58) becomes

$$\begin{aligned} s_{ssb}(t)\cos 2\pi(f_c + \Delta f)t = &\\ \frac{s(t)\cos 2\pi\Delta f t}{4} + &\frac{s(t)\cos 4\pi(f_c + \Delta f)t}{4} \\ + \frac{\hat{s}(t)\sin 2\pi\Delta f t}{4} + &\frac{\hat{s}(t)\sin 4\pi(f_c + \Delta f)}{4} \end{aligned} \tag{6.61}$$

It is clear that only the first and third terms go through the lowpass filter. If only the first term appeared in the output, the effect would be the same as that experienced in double sideband. However, the addition of the Hilbert transform term changes the output. The specific form of that change can be seen if we consider the special case of a sinusoidal modulating signal. That is, let $s(t) = \cos 2\pi f_m t$. The output of the lowpass filter is then proportional to $\cos 2\pi(f_m \pm \Delta f)t$, with the plus sign obtaining for lower sideband and the minus sign for upper sideband. (You can prove this either from Eq. (6.61) or by writing the single-sideband waveform as a single sinusoid at a frequency of either $f_c - f_m$ or $f_c + f_m$, depending upon whether lower or upper sideband is being considered.) The effect of the frequency mismatch is therefore a frequency offset in the demodulated wave. If the information signal is a sum of sinusoids, the demodulated signal will be a sum of shifted sinusoids. Thus, in the general case, we see an overall shifting of frequencies of $s(t)$ by the amount of the frequency mismatch. You might think that this results in a change in pitch of the sound, but such is not the case. For example, suppose you hummed into a microphone. Then the resulting waveform would be periodic with a fundamental frequency equal to the frequency at which you are humming. Harmonics would occur at multiples of this frequency. If each frequency component is shifted by the same amount, the relationship among the harmonics is destroyed, and the resulting sound changes. The effect upon music is generally considered to be unacceptable. The effect upon voice is sometimes acceptable, since the resulting sound is usually intelligible (at least, for small Δf relative to the frequencies that are present). Some have described the effect as a "Donald Duck" voice.

Therefore, while mismatches are to be avoided, the effects may be considered less devastating than in the case of double sideband.

With double sideband, we made demodulation simpler by adding a carrier. We can also add a carrier to the single-sideband waveform, recognizing that the carrier would be at the edge of the band of frequencies occupied by the waveform. The carrier could be extracted using a filter at the receiver, or a phase lock loop could be used.

Can we use an incoherent detector, such as the envelope detector? To answer this question, let us examine transmitted carrier lower sideband, which is given by

$$s_{\text{lsb}}(t) + A\cos 2\pi f_c t = \left[A + \frac{s(t)}{2}\right]\cos 2\pi f_c t + \frac{\hat{s}(t)\sin 2\pi f_c t}{2} \tag{6.62}$$

The envelope of this waveform is found by combining the cosine and sine into a single sinusoid with a time-varying amplitude and phase. The envelope is given by

$$\sqrt{\left[A + \frac{1}{2}s(t)\right]^2 + \left[\frac{1}{2}\hat{s}(t)\right]^2} \tag{6.63}$$

In general, this does not look anything like $s(t)$. However, if the constant A is very large, the first term predominates, and the expression is approximately equal to $A + s(t)/2$. Of course, large values of A mean inefficient operation, with most of the transmitted energy going into the carrier. Therefore, a system of this type finds limited application.

6.9 VESTIGIAL SIDEBAND

The advantage of single over double sideband (SSB and DSB) is the former's economy of frequency usage. That is, SSB uses half the corresponding bandwidth required for DSB transmission. The primary disadvantage of SSB is the difficulty in building a transmitter or an effective receiver for it. One problem is that when we attempt to build a sharp filter to remove one of the sidebands, the phase characteristic of the filter develops ripple. The closer one approaches the amplitude characteristic of the ideal filter, the worse becomes the phase characteristic. One area in which frequency conservation becomes critical is television, where bandwidths are orders of magnitude greater than those used for voice transmission. Phase distortion in a video signal causes offset of the resulting scanned image, and this is seen as ghost images on the screen. The eye is much more sensitive to such forms of distortion than is the ear to equivalent forms of voice distortion. We therefore have reason to explore a compromise between SSB and DSB.

Vestigial sideband (VSB) possesses a frequency bandwidth advantage approaching that of single sideband without the disadvantage of difficulty in building a modulator. It is also easier to construct a demodulator for this form of communication.

As the name implies, VSB includes a vestige, or trace, of the second sideband. Thus, instead of completely eliminating the second sideband, as in the case of SSB, we eliminate most, but not all, of it.

Suppose we begin with DSB but filter out one of the sidebands. In contrast to SSB, with VSB we use a filter that does *not* closely approach the ideal infinite roll-off. The re-

sult might resemble Fig. 6.45, where we show the double-sided transform, the filter characteristic, and the output transform that is generated.

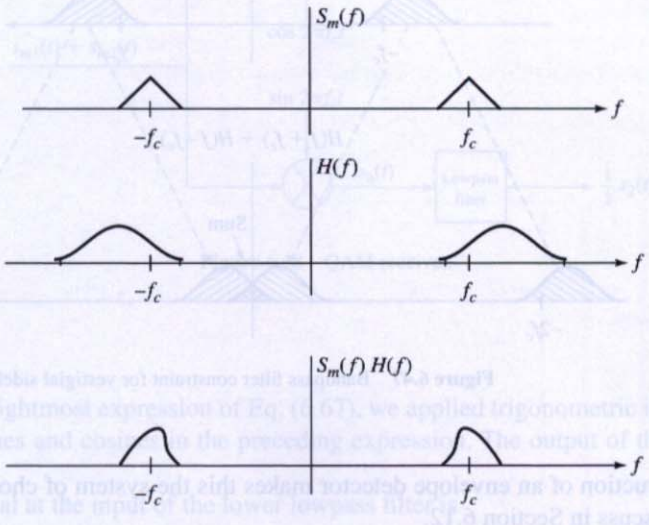


Figure 6.46 Vestigial sideband generator.

Suppose now that this output signal forms the input to a synchronous (coherent) demodulator. Then, when the VSB signal is multiplied by a cosine at the carrier frequency, the Fourier transform shifts both up and down by the carrier frequency. The part that shifts down passes through the lowpass filter. The part that shifts up resides around $2f_c$ and is rejected by the filter. The filter output then has a transform given by

$$S_0(f) = \frac{S(f)[H(f + f_c) + H(f - f_c)]}{4} \tag{6.64}$$

Equation (6.64) can be used to set the conditions on the filter. The bracketed sum is shown in Fig. 6.47 for a typical filter transfer function $H(f)$. This filter transfer function, $H(f)$, must display odd symmetry for frequencies around the carrier such that the sum of the two terms approximates a constant characteristic. The tail of the filter characteristic must be asymmetric about $f = f_c$. That is, the output half of the tail must fold over and fill in any difference between the inner-half values and the value for an ideal filter.

Suppose that we add a carrier term to a vestigial sideband signal. The vestigial sideband transmitted carrier waveform is then of the form

$$s_v(t) + A \cos 2\pi f_c t \tag{6.65}$$

This carrier term can be extracted at the receiver using either a very narrow bandpass filter or a phase lock loop. If the carrier term is large enough, an incoherent detector (e.g., an envelope detector) can be used. We said this same thing in regard to SSB, where the carrier had to be much larger than the signal. In DSB, the carrier need only be of the same order of magnitude as the signal. The required carrier size for VSB is between these two extremes. While the addition of a strong carrier significantly decreases efficiency, the ease of

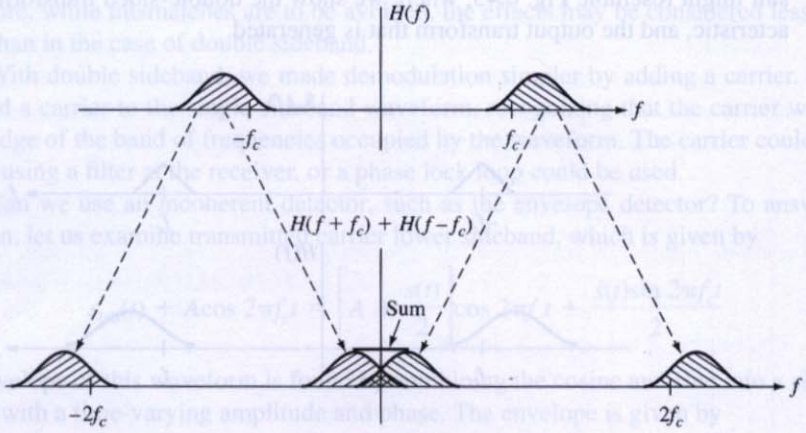


Figure 6.47 Bandpass filter constraint for vestigial sideband.

construction of an envelope detector makes this the system of choice in television, which we discuss in Section 6.12.

6.10 HYBRID SYSTEMS AND AM STEREO

We now investigate a hybrid modulation technique that permits a new form of multiplexing of signals. We have repeatedly stated that signals can be separated, provided that they do not overlap in time or in frequency. Double-sideband AM maintains frequency separation in order to keep channels from interfering with each other, but it uses twice the bandwidth of single sideband. The latter fact, however, hints that it might be possible to send two double-sideband AM signals that overlap in both time and frequency *and yet still be able to separate them* at the receiver. Indeed, *quadrature amplitude modulation* (QAM) accomplishes this.

Suppose we have two information signals $s_1(t)$ and $s_2(t)$, each of which is limited to frequencies below f_m. We now modulate two carriers of exactly the same frequency with these two signals. However, the carriers are in phase quadrature (90° out of phase) with each other. The sum of the two AM waveforms is then

$$s_{m1}(t) + s_{m2}(t) = s_1(t)\cos 2\pi f_c t + s_2(t)\sin 2\pi f_c t \tag{6.66}$$

Even though the two AM waveforms overlap in both frequency and time, they can be separated by the receiver shown in Fig. 6.48. The signal at the input of the upper lowpass filter is

$$\begin{aligned} s_a(t) &= s_1(t)\cos^2 2\pi f_c t + s_2(t)\sin 2\pi f_c t\cos 2\pi f_c t \\ &= \frac{1}{2}[s_1(t) + s_1(t)\cos 4\pi f_c t + s_2(t)\sin 4\pi f_c t] \end{aligned} \tag{6.67}$$

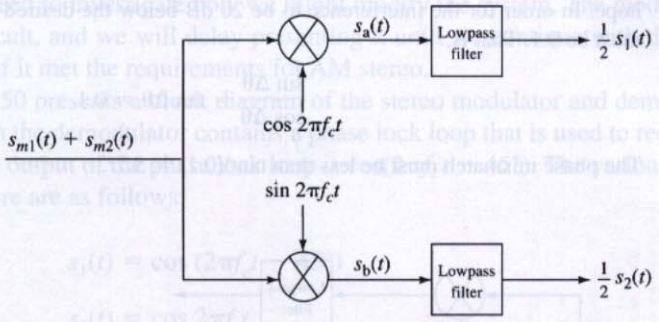


Figure 6.48 QAM receiver.

To derive the rightmost expression of Eq. (6.67), we applied trigonometric identities to the products of sines and cosines in the preceding expression. The output of the filter is then $s_1(t)/2$.

The signal at the input of the lower lowpass filter is

$$s_b(t) = s_1(t)\cos 2\pi f_c t \sin 2\pi f_c t + s_2(t)\sin^2 2\pi f_c t \tag{6.68}$$

$$= \frac{1}{2}[s_1(t)\sin 4\pi f_c t + s_2(t) - s_2(t)\cos 4\pi f_c t]$$

The signal at the output of the filter is then $s_2(t)/2$, and separation is accomplished. Of course, this scheme requires perfect phase control at the receiver to avoid having one signal interfere with the other. Hence, since phase carries some of the information, incoherent detection cannot be used; we must be able to recover the carrier precisely.

Example 6.6

A QAM scheme of the type shown in Fig. 6.48 is used to simultaneously transmit two waveforms in a channel in the frequency range $f_c \pm f_m$. The oscillators in the receiver are in error by $\Delta\theta$. Assuming that the information signals are sinusoids of equal amplitude, find the maximum value of $\Delta\theta$ such that the interference (cross talk) is limited to -20 dB.

Solution: We rederive Eqs. (6.66) and (6.67) for the receiver shown in Fig. 6.49. The output of the upper filter is now

$$\frac{1}{2}[s_1(t)\cos\Delta\theta + s_2(t)\sin\Delta\theta]$$

and the output of the lower filter is

$$\frac{1}{2}[-s_1(t)\sin\Delta\theta - s_2(t)\cos\Delta\theta]$$

The ratio of the amplitude of the undesired term to that of the desired signal is $\sin\Delta\theta/\cos\Delta\theta$. Note that as the phase error approaches zero, this ratio also approaches zero, as we would

hope. In order for the interference to be 20 dB below the desired signal, the amplitude ratio must be 0.1. That is,

$$\frac{\sin \Delta\theta}{\cos \Delta\theta} = \tan \Delta\theta = 0.1$$

The phase mismatch must be less than $\tan^{-1}(0.1) = 5.7°$.

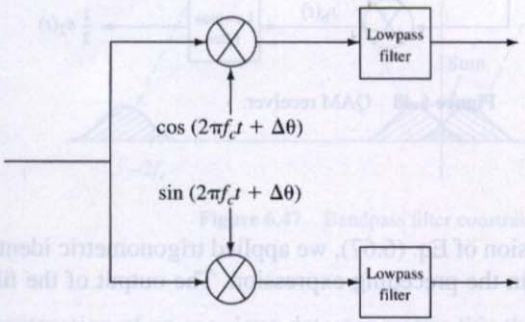


Figure 6.49 QAM receiver with phase mismatch.

AM Stereo

The idea behind AM stereo is to send two independent audio signals within the 10-kHz bandwidth allocated by the FCC to each commercial broadcast station. Additionally, the FCC requires compatibility with monaural receivers. Thus, if the two audio signals represent the left and right channels of a stereo system, a monaural receiver must recover the sum of these two signals.

Quadrature amplitude modulation represents one technique to send two signals simultaneously. If the two signals are designated as $s_L(t)$ and $s_R(t)$, the composite signal can be written as

$$q(t) = s_L(t)\cos 2\pi f_c t + s_R(t)\sin 2\pi f_c t \tag{6.69}$$

If both $s_L(t)$ and $s_R(t)$ are audio signals with a maximum frequency of 5 kHz, $q(t)$ occupies the band of frequencies between $f_c - 5$ kHz and $f_c + 5$ kHz, a total bandwidth of 10 kHz. The composite signal can then be rewritten as the single sinusoid

$$q(t) = A(t)\cos\left[2\pi f_c t + \theta(t)\right] \tag{6.70a}$$

where

$$\begin{aligned} A(t) &= \sqrt{s_L^2(t) + s_R^2(t)} \\ \theta(t) &= -\tan^{-1}\left(\frac{s_R(t)}{s_L(t)}\right) \end{aligned} \tag{6.70b}$$

The envelope detector in a monaural AM receiver would produce $A(t)$. This is a distorted version of the sum of the two channels and does not meet the compatibility requirement.

We therefore need to investigate how we might modify the system. The modification does not prove difficult, and we will delay presenting it until we continue with the analysis of this system as if it met the requirements for AM stereo.

Figure 6.50 presents a block diagram of the stereo modulator and demodulator. The dashed block in the demodulator contains a phase lock loop that is used to recover the carrier. In fact, the output of the phase lock loop is $\cos(2\pi f_c t - 45°)$. The various functions of time in the figure are as follows:

$$
\begin{aligned}
s_1(t) &= \cos(2\pi f_c t - 45°) \\
s_2(t) &= \cos 2\pi f_c t \\
s_3(t) &= \sin 2\pi f_c t \\
s_4(t) &= s_L(t)\cos^2 2\pi f_c t + s_R(t)\sin 2\pi f_c t \cos 2\pi f_c t \\
s_5(t) &= s_L(t)\sin 2\pi f_c t \cos 2\pi f_c t + s_R(t)\sin^2 2\pi f_c t \\
s_6(t) &= \frac{1}{2} s_L(t) \\
s_7(t) &= \frac{1}{2} s_R(t)
\end{aligned}
\tag{6.71}
$$

Now that we see that the two channels can be separated, we present several specific techniques which assure compatibility. One technique (proposed in two different forms, by Belar and by Magnavox) angle modulates the carrier with one audio signal and amplitude modulates the resulting modulated carrier with the second signal. The angle modulation is narrowband. (We discuss angle modulation in the next chapter.)

Figure 6.51 presents one possible AM/FM configuration: a simplified version of the Belar AM stereo system. The frequency deviation of the FM is $\Delta f = 320$ Hz, which is much less than the maximum audio frequency of 5 kHz. This assures that the resulting modulated signal can be kept within the 10-kHz assigned band. The system meets the compatibility requirement, since an envelope detector will recover the sum signal. The limiter in the receiver removes the amplitude modulation, leaving an FM wave with approximately constant amplitude.

The idealized system of Fig. 6.51 requires some modification to make it practical. Timing is important in a stereo system. If the L − R and $L + R$ channels are not synchronized, the original signals cannot be recovered without distortion. Since the paths traversed by the L − R and $L + R$ signals do not take identical lengths of time, it is necessary to insert a time delay in one of the lines so that the two signals can be properly aligned at the output.

A more complex technique of stereo transmission starts by amplitude modulating a frequency-modulated carrier with the sum signal, $L + R$, as in the preceding system. However, the carrier is not angle modulated with the difference signal. Instead, the angle modulation is performed in such a way that the information in the left channel is carried on the lower sideband and that in the right channel on the upper sideband. To do so requires a relatively complex system involving a 90° phase shift between $L + R$ and $L - R$ signals and also employing automatic gain control (AGC) circuitry. One advantage of this system is

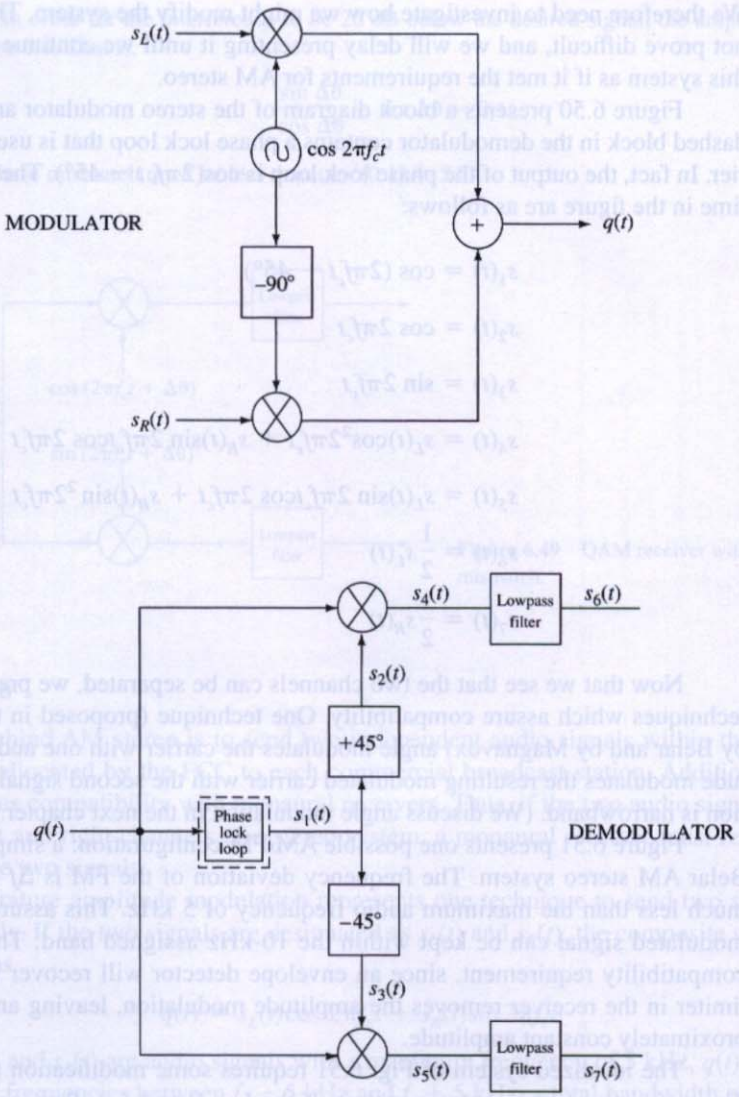


Figure 6.50 Quadrature modulation stereo system.

that it is possible to produce a stereo effect with two monaural receivers merely by tuning one receiver slightly above, and the other slightly below, the carrier.

Yet a third technique is a refinement of the QAM stereo system discussed at the beginning of this section. Recall that the problem with that system was one of compatibility: A monaural receiver would recover the square root of the sum of the squares of the left and right signals, a distorted version of the sum waveform.

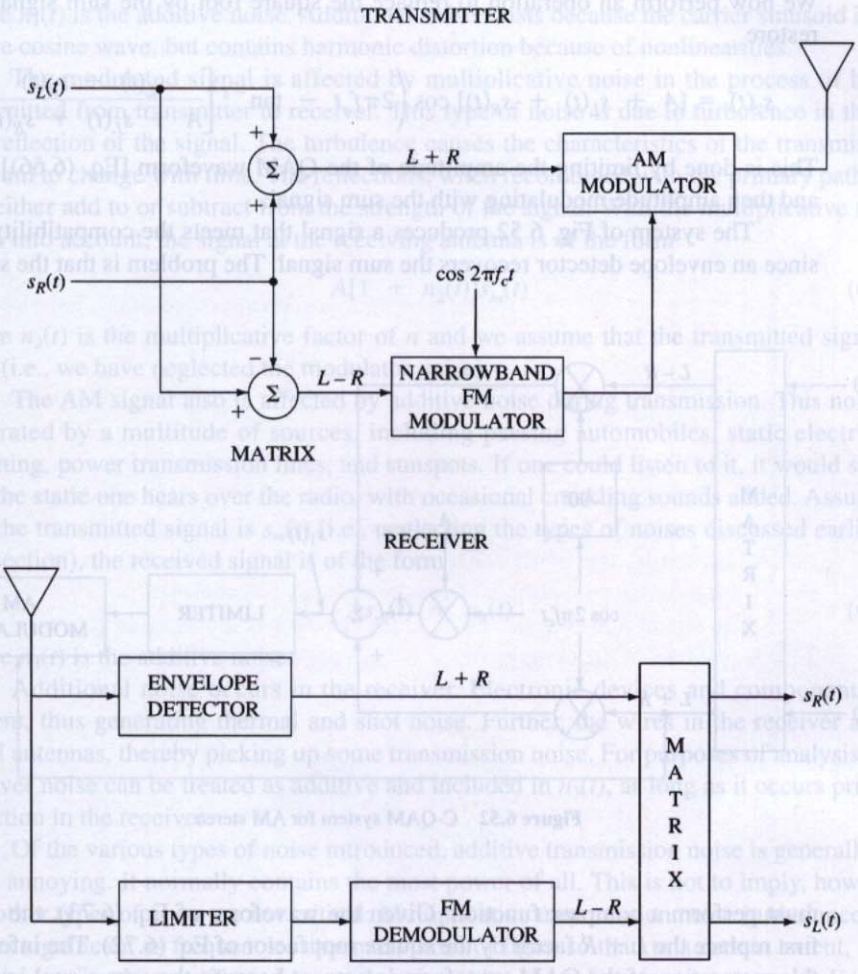


Figure 6.51 AM/FM system for AM stereo.

A *compatible* QAM (C-QAM) system (a simplified version of the Motorola configuration) is shown in Fig. 6.52. The system begins with a QAM signal whose two components are the sum and difference waveforms. The signal at $s_1(t)$ is

$$s_1(t) = \sqrt{[a + s_L(t) + s_R(t)]^2 + [s_L(t) - s_R(t)]^2} \times \cos\left(2\pi f_c t - \tan^{-1}\left[\frac{s_L(t) - s_R(t)}{A + s_L(t) + s_R(t)}\right]\right) \tag{6.72}$$

We now perform an operation to replace the square root by the sum signal. That is, we restore

$$s_2(t) = [A + s_L(t) + s_R(t)] \cos\left(2\pi f_c t - \tan^{-1}\left[\frac{s_L(t) - s_R(t)}{A + s_L(t) + s_R(t)}\right]\right) \quad (6.73)$$

This is done by limiting the amplitude of the QAM waveform [Eq. (6.66)] to a constant and then amplitude modulating with the sum signal.

The system of Fig. 6.52 produces a signal that meets the compatibility requirement since an envelope detector recovers the sum signal. The problem is that the stereo receiver

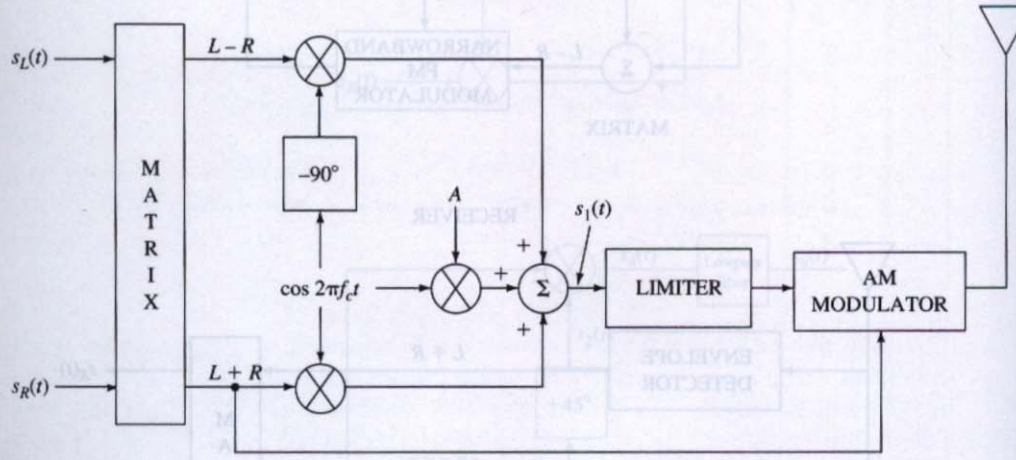


Figure 6.52 C-QAM system for AM stereo.

must perform a complex function. Given the waveform of Eq. (6.73), the receiver must first replace the $L + R$ factor by the square root factor of Eq. (6.72). The information to do this restoration of the QAM waveform is present because the sum signal exists as the envelope and the difference signal can be found by detecting the phase and combining it with the sum signal waveform. The problem is that the receiver must perform these operations with relatively simple circuitry. The receiver uses a phase lock loop to detect the phase and then remodulates the received waveform with that phase.

6.11 PERFORMANCE

In the process of communicating a signal, noise arises in various ways. The information signal $s(t)$ is corrupted by some noise before it even reaches the modulator in the transmitter. This noise is generated by electronic devices in the modulator. Thus, the signal at the output of the modulator is of the form

$$[s(t) + n_1(t)]\cos 2\pi f_c t \quad (6.74)$$

where $n_1(t)$ is the additive noise. Additional noise exists because the carrier sinusoid is not a pure cosine wave, but contains harmonic distortion because of nonlinearities.

The modulated signal is affected by multiplicative noise in the process of being transmitted from transmitter to receiver. This type of noise is due to turbulence in the air and reflection of the signal. The turbulence causes the characteristics of the transmission medium to change with time. The reflections, when recombined with the primary path signal, either add to or subtract from the strength of the signal. With the multiplicative noise taken into account, the signal at the receiving antenna is of the form

$$A[1 + n_2(t)]s_m(t) \tag{6.75}$$

where $n_2(t)$ is the multiplicative factor of n and we assume that the transmitted signal is $s_m(t)$ (i.e., we have neglected the modulator noise).

The AM signal also is affected by additive noise during transmission. This noise is generated by a multitude of sources, including passing automobiles, static electricity, lightning, power transmission lines, and sunspots. If one could listen to it, it would sound like the static one hears over the radio, with occasional crackling sounds added. Assuming that the transmitted signal is $s_m(t)$ (i.e., neglecting the types of noises discussed earlier in this section), the received signal is of the form

$$As_m(t) + n_3(t) \tag{6.76}$$

where $n_3(t)$ is the additive noise.

Additional noise occurs in the receiver. Electronic devices and components are present, thus generating thermal and shot noise. Further, the wires in the receiver act as small antennas, thereby picking up some transmission noise. For purposes of analysis, this receiver noise can be treated as additive and included in $n_3(t)$, as long as it occurs prior to detection in the receiver.

Of the various types of noise introduced, additive transmission noise is generally the most annoying. It normally contains the most power of all. This is not to imply, however, that other types of noise are not critical: Multiplicative transmission noise (turbulence) becomes significant as frequencies approach those of light, so that, to a certain extent, UHF television signals are affected by such noise. Also, very low-frequency multiplicative noise causes fading in microwave systems.

6.11.1 Coherent Detection

Double Sideband

We first examine the case of double-sideband suppressed carrier transmission and synchronous demodulation. We assume that we have been able to match the carrier frequency and phase exactly.

The received waveform at the input to the receiver is

$$r(t) = Ks(t)\cos 2\pi f_c t + n(t) \tag{6.77}$$

where K is a constant that accounts for attenuation during transmission and $n(t)$ is the additive noise. We assume that $n(t)$ is white Gaussian noise with two-sided power spectral density $N_0/2$. That is, the noise power is N_0 watts/Hz.

Let us begin by finding the signal to noise ratio at the *input* to the synchronous demodulator of Fig. 6.53. The synchronous demodulator contains a bandpass filter that did not appear in our earlier discussions. This filter is known as a *predetection filter*. In a theoretical analysis it is redundant, since any frequencies rejected by it would also be rejected

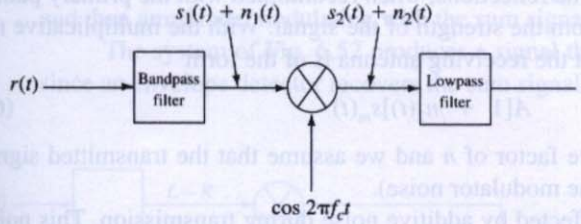


Figure 6.53 Synchronous demodulator.

by the final lowpass filter. It is included both for practical reasons and to simplify the analysis. The electronic device that performs the multiplication could get overloaded and driven into saturation if the input contained too much energy outside of the band of interest. Additionally, the white noise at the input to the system contains (ideally) infinite power. Including it as well would also complicate the analysis.

The signal power at the input to the detector is the average power of $Ks(t)\cos 2\pi f_c t$. This is one half of the average of the square of the cosine amplitude, or

$$\frac{\overline{K^2 s^2(t)}}{2} = \frac{K^2 P_s}{2} \tag{6.78}$$

where P_s is the average power of $s(t)$. The average power of the filtered noise is N_o times the bandwidth of the filter, or $2f_m N_o$. The input signal-to-noise ratio, SNR_i, is then

$$\text{SNR}_i = \frac{K^2 P_s}{4 N_0 f_m} \tag{6.79}$$

We now wish to derive the signal to noise ratio at the output. The signal at the output of the synchronous demodulator is $Ks(t)/2$, and its average power is $K^2 P_s/4$. To find the noise at the output of the detector, we must turn our attention to the time domain. That is, since the demodulator performs the non-linear operation of multiplication, we can no longer track noise power through the system using the power spectral density.

The bandlimited noise at the detector input can be expanded into its quadrature components thus:

$$n_1(t) = x(t)\cos 2\pi f_c t - y(t)\sin 2\pi f_c t \tag{6.80}$$

The power spectral densities of $x(t)$ and $y(t)$ are calculated as in the example in the Section 4.5. These are shown in Fig. 6.54. The noise at the input to the lowpass filter is

$$\begin{aligned} n_2(t) &= [x(t)\cos 2\pi f_c t - y(t)\sin 2\pi f_c t]\cos 2\pi f_c t \\ &= \frac{x(t) + x(t)\cos 4\pi f_c t - y(t)\sin 4\pi f_c t}{2} \end{aligned} \tag{6.81}$$

where we have used the trigonometric identities

$$\cos^2 2\pi f_c t = \frac{1}{2} + \frac{1}{2}\cos 4\pi f_c t$$

and (6.82)

$$\sin 2\pi f_c t \cos 2\pi f_c t = \frac{1}{2}\sin 4\pi f_c t$$

to obtain the rightmost expression in Eq. (6.81).

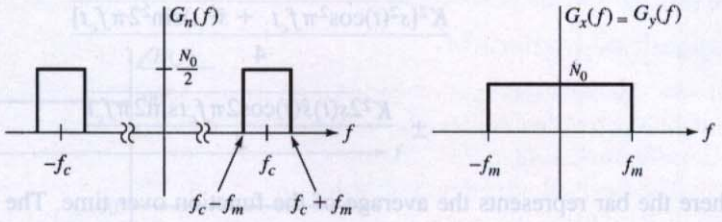


Figure 6.54 Quadrature expansion of noise.

The only term in Eq. (6.81) that passes through the lowpass filter is the first term, so the noise at the output of the detector is $x(t)/2$. The power of this is the power of $x(t)$ divided by 4. The power of $x(t)$ is found by integrating $G_x(f)$, so the total noise power at the detector output is $N_0 f_m/2$. We found the signal power to be $K^2 P_s/4$, so the output signal-to-noise ratio is

$$\text{SNR}_0 = \frac{K^2 P_s}{2 f_m N_0} \tag{6.83}$$

Comparing this to Eq. (6.79), we find that

$$\text{SNR}_0 = 2\text{SNR}_i \tag{6.84}$$

The demodulation process has doubled the signal to noise ratio. Let us try to give an intuitive justification of this result. In double sideband, the two sidebands are related to each other. Thus, knowing one of the sidebands, you can derive the second. The synchronous demodulator essentially realigns the two sidebands to add to each other, so the effect upon the signal is similar to adding a signal to itself. This *coherent* addition doubles the amplitude and therefore multiplies the power by a *factor of four*. On the other hand, the noises in the two sidebands are unrelated (i.e., independent). When these two noise sources are added, it is like adding two independent noise processes. In that case, the mean square values add, and the power *doubles*. The signal power has increased by a factor of four, while the noise power has only doubled. Therefore, the signal to noise ratio doubles.

Single Sideband

We now repeat the foregoing analysis for single sideband. The received signal is of the form

$$r(t) = \frac{Ks(t)\cos 2\pi f_c t \pm K\hat{s}(t)\sin 2\pi f_c t}{2} + n(t) \tag{6.85}$$

where, once again, K is a constant that accounts for attenuation during transmission and $n(t)$ is the noise at the output of the predetection filter of Fig. 6.53.

The signal power at the input to the detector is the average of the square of the signal portion of $r(t)$. This is equal to

$$\frac{K^2[\overline{s^2(t)\cos^2\pi f_c t} + \overline{\hat{s}^2(t)\sin^2 2\pi f_c t}]}{4} \pm \frac{K^2\overline{2s(t)\hat{s}(t)\cos 2\pi f_c t \sin 2\pi f_c t}}{4} \tag{6.86}$$

where the bar represents the average of the function over time. The last term is equal to zero, since the average of the cosine multiplied by the sine is zero. (Take the integral over one period.) The squares of the sinusoids have an average value of $\frac{1}{2}$, so the input signal power becomes

$$P = \frac{K^2(P_s + P_{\hat{s}})}{8} \tag{6.87}$$

where P_s is the power of $s(t)$ and $P_{\hat{s}}$ is the power of the Hilbert transform $\hat{s}(t)$. The Hilbert transform results from putting $s(t)$ through a filter with $H(f) = -j\text{sgn}(f)$, as shown in Fig. 6.55.

The output power spectral density is given by

$$G_{\hat{s}}(f) = |H(f)|^2 G_s(f) \tag{6.88}$$

The square of the magnitude of the filter characteristic is unity, so the power of the Hilbert transform is the same as the power of the original signal. Therefore, the input signal power is found from Eq. (6.87) to be $K^2P_s/4$. The input noise power is N_0 multiplied by the bandwidth of the bandpass filter, or $N_0 f_m$. The signal to noise ratio at the detector input is then

$$\text{SNR}_i = \frac{K^2 P_s}{4N_0 f_m} \tag{6.89}$$

We now turn our attention to the detector output. The signal at the output is given by $Ks(t)/4$. If we expand the detector input noise in a quadrature expansion, as in Eq. (6.80) for double sideband, the output noise is again given by $x(t)/2$. However, the power spectral density of $x(t)$ is not the same as that given in Fig. 6.54. Figure 6.56 shows the power spectral density of $x(t)$.

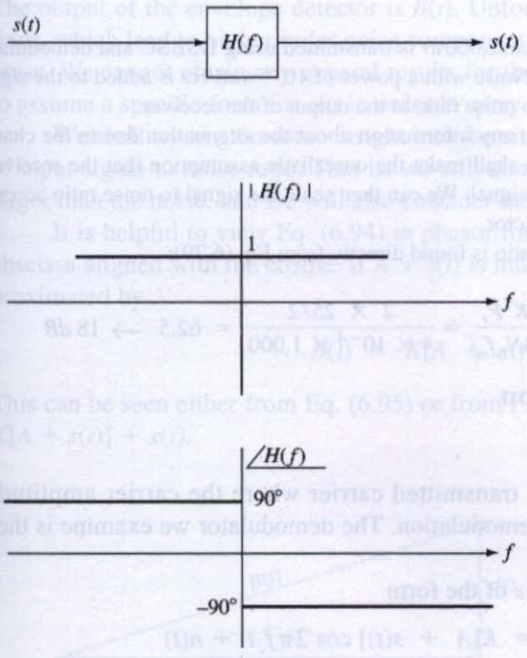


Figure 6.55 The Hilbert transform.

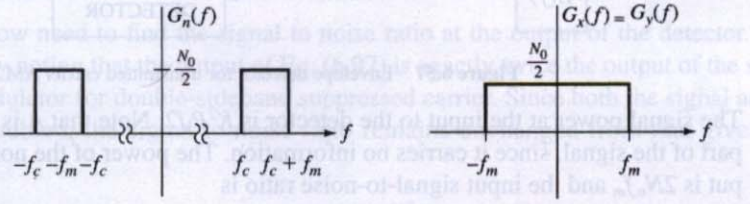


Figure 6.56 Power spectral density of quadrature noise.

The output signal power is $K^2P_s/16$, and the output noise power is $N_0f_m/4$. The output signal to noise ratio is then

$$\text{SNR}_0 = \frac{K^2P_s}{4N_0f_m} \tag{6.90}$$

which *is the same* as the signal to noise ratio at the detector input. That is,

$$\text{SNR}_0 = \text{SNR}_i \tag{6.91}$$

This result distinguishes single sideband from double sideband. Indeed, one would expect to get some benefit from using twice the bandwidth.

Example 6.7

A baseband signal $s(t) = 5\cos 2{,}000\pi t$ is transmitted using DSBSC and demodulated using a synchronous demodulator. Noise with a power of 10^{-4} watt/Hz is added to the signal prior to reception. Find the signal to noise ratio at the output of the receiver.

Solution: We do not have any information about the attenuation due to the channel or due to the antenna patterns. We shall make the unrealistic assumption that the received signal is identical to the transmitted signal. We can then scale the signal to noise ratio according to the square of any attenuation factor.

The signal to noise ratio is found directly from Eq. (6.79):

$$\text{SNR}_0 = \frac{K^2 P_s}{4N_0 f_m} = \frac{2 \times 25/2}{4 \times 10^{-4} \times 1{,}000} = 62.5 \longrightarrow 18\ dB$$

6.11.2 Incoherent Detection

Transmitted Carrier AM

We now consider the case of transmitted carrier where the carrier amplitude is large enough to permit incoherent demodulation. The demodulator we examine is the envelope detector shown in Fig. 6.57.

The received waveform is of the form

$$r(t) = K[A + s(t)] \cos 2\pi f_c t + n(t) \tag{6.92}$$

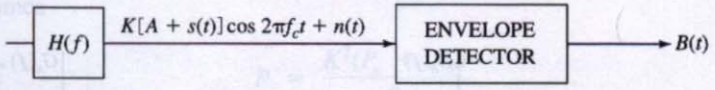


Figure 6.57 Envelope detector for transmitted carrier AM.

The signal power at the input to the detector is $K^2P_s/2$. Note that A is not considered to be part of the signal, since it carries no information. The power of the noise at the detector input is $2N_0 f_m$ and the input signal-to-noise ratio is

$$\text{SNR}_i = \frac{K^2 P_s}{4N_0 f_m} \tag{6.93}$$

Note that this is identical to the input signal-to-noise ratio for double-sideband *synchronous* demodulation [Eq. (6.79)].

To find the output of the detector, we expand the input noise into quadrature form and then combine terms into a single sinusoid. We obtain

$$\begin{aligned} &[KA + Ks(t) + x(t)]\cos 2\pi f_c t - y(t)\sin 2\pi f_c t \\ &\quad = B(t)\cos[2\pi f_c t + \theta(t)] \end{aligned} \tag{6.94}$$

where

$$B(t) = \sqrt{[KA + Ks(t) + x(t)]^2 + y^2(t)} \tag{6.95}$$

and

$$\theta(t) = -\tan^{-1}\left(\frac{y(t)}{KA + Ks(t) + x(t)}\right) \tag{6.96}$$

The output of the envelope detector is $B(t)$. Unfortunately, this contains nonlinear operations, which lead to higher order noise components and cross products between signal and noise. We cannot obtain any general results for the output signal to noise and would have to assume a specific form for $s(t)$ in order to carry the analysis further.

We shall try to gain some insight into the situation by considering the limiting cases of input signal to noise ratio. That is, we will consider the case where the signal is much larger than the noise, and we will also consider the opposite situation.

It is helpful to view Eq. (6.94) in phasor form. Figure 6.58 illustrates this, with the abscissa aligned with the cosine. If $A + s(t)$ is much larger than the noise, $B(t)$ can be approximated by

$$B(t) \approx K[A + s(t)] + x(t) \tag{6.97}$$

This can be seen either from Eq. (6.95) or from Fig. 6.58, where we assume that $y(t) << K[A + s(t)] + x(t)$.

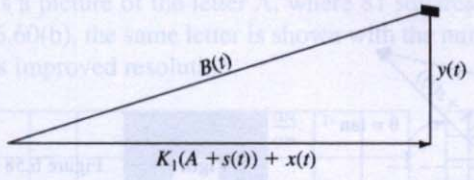


Figure 6.58 Phasor diagram of envelope detector input.

We now need to find the signal to noise ratio at the output of the detector. We can save time by noting that the output of Eq. (6.97) is exactly twice the output of the synchronous demodulator for double-sideband suppressed carrier. Since both the signal and noise ratio are doubled, the signal to noise ratio remains unchanged from that given in Eq. (6.83):

$$\mathrm{SNR}_0 = \frac{K^2 P_s}{2 f_m N_0} \tag{6.98}$$

The ratio of output SNR to input SNR is therefore the same as for double-sideband coherent demodulation. That is,

$$\mathrm{SNR}_0 = 2\mathrm{SNR}_i \tag{6.99}$$

This identical result can be deceiving. We must bear in mind that the price we are paying is lower efficiency. In comparing the various systems, it is important to do so under equivalent conditions. We can get an approximate result by assuming that $s(t)$ is a pure sinusoid, $\cos 2\pi f_m t$. Since this sinusoid has a maximum negative excursion of -1, the minimum value of A is 1. Using this value, we find that $A + s(t)$ is given by

$$1 + \cos 2\pi f_m t \tag{6.100}$$

The signal power at the output is $\frac{1}{2}$ watt, and the power of the dc term in the output is 1 watt. Thus, the true signal to noise ratio at the output is one-third of that found in Eq.

(6.98). In comparing systems, we therefore often use the following expression for incoherent detection:

$$\mathrm{SNR}_0 = \frac{2\mathrm{SNR}_i}{3} \tag{6.101}$$

We now consider the other extreme in incoherent detection, that is, a very low signal to noise ratio. To analyze this situation, we will redraw Fig. 6.58, but this time referenced to the noise signal. That is, we add the signal vector to the larger noise vector. The realigned diagram is shown in Fig. 6.59. Note that the angle between the signal and noise is

$$\theta(t) = \tan^{-1}\left(\frac{y(t)}{x(t)}\right) \tag{6.102}$$

The resultant vector is approximately given by

$$\sqrt{x^2(t) + y^2(t)} + K[A + s(t)]\cos\theta(t) \tag{6.103}$$

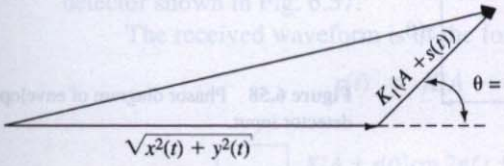


Figure 6.59 Figure 6.58 redrawn for low SNR.

The only place where the signal appears is in the last term, and it is multiplied by a random noise term, $\cos\theta(t)$. Thus, we have *both additive and multiplicative noise*. It can be shown that $\theta(t)$ is uniformly distributed between 0° and 360°. It should therefore not be surprising to observe that the signal *cannot* be recovered from the envelope detector output.

As the signal to noise ratio decreases from a high value, a threshold is reached. For signal to noise ratios above this threshold, the output signal to noise ratio is linearly related to the detector input signal to noise ratio. For signal to noise ratios below this threshold, the dependence approaches a quadratic relationship. That is, for each decrease in input signal to noise ratio by a factor of 2, the output signal to noise ratio decreases approximately by a factor of 4.

6.12 TELEVISION

Public television had its beginnings in England in 1927. In the United States, it started three years later, in 1930. These early forms used *mechanical* scanning of the picture to be transmitted. That is, a picture was changed into an electrical signal by scanning the entire image along a spiral starting at the center. The scanning was accomplished by means of a rapidly rotating wheel with holes cut in it. As the wheel rotated, light from various parts of the total picture passed through the holes.

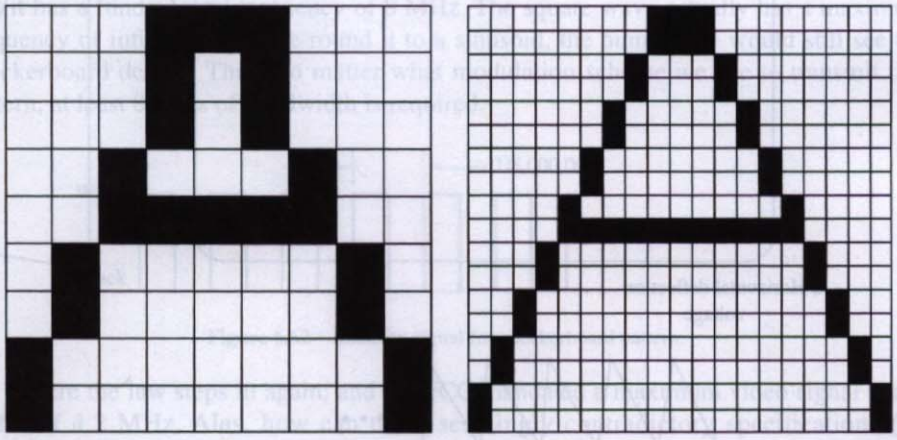
During this early period, broadcasts did not follow any regular schedule. Such regular scheduling did not begin until 1939, during the opening of the New York World's Fair.

The concepts of television and picture transmission spread into many exciting areas; facsimile transmission, satellite broadcasts, video telephone, video-text, and cable TV represent only a few examples. A cable TV revolution is occurring, with two-way communication links becoming common. We can anticipate that TV will eventually replace the newspaper, supermarket, baby-sitter, theater, and perhaps (heaven forbid!) the university campus. The theory to be presented here, although geared toward broadcast TV, is applicable to most forms of picture transmission.

A Picture Is Worth A Thousand Words?

Nonsense! A picture is equivalent to far more than a thousand words. While we could rigorously define the information content of a picture using concepts from the science of information theory, we will not do that here. Instead, we will subdivide the picture into small components similar to words.

Suppose we divide a picture into squares, where each square is a certain shade. The number of squares in any given area determines the *resolution*. For example, Fig. 6.60(a) shows a picture of the letter A, where 81 squares have been used to define the picture. In Fig. 6.60(b), the same letter is shown with the number of squares increased to 342. The result is improved resolution.



(a) (b)

Figure 6.60 Definition of picture elements.

The number of squares used in U.S. television is set by the FCC at 211,000 picture elements (*pixels*). This is divided into 426 elements in each horizontal line and 495 visible horizontal lines in each picture. The job of television transmitters is then to step through the 211,000 picture elements and send an intensity value for each one. The receiver interprets these transmissions and reconstructs the picture from the 211,000 intensity levels. This information must be updated rapidly to simulate motion. (We confine our attention here to monochrome (black-and-white) television.)

A conventional TV receiver is not much different from a conventional analog laboratory oscilloscope. A beam of electrons is shot toward a screen and bent by deflection plates. When a negative charge is placed on a plate, the electron beam is repelled. In the

oscilloscope, we apply a sawtooth waveform to the horizontal deflection plates to sweep the beam from left to right (creating a time axis) and then more rapidly back to the left. This traces a line. In TV receivers, we add a second dimension to the sweep. While the beam is rapidly sweeping from left to right on the screen, it is less rapidly sweeping from top to bottom. The net result is a series of almost horizontal lines on the screen, as sketched in Fig. 6.61. This is known as the TV *raster*. We are describing traditional analog scanning TV. Digital television has the option of controlling the choice of picture elements using digital counting circuitry. Picture elements are not necessarily bombarded with an aimed electron beam, but could be implemented as individual LED or LCD elements arranged in a matrix and digitally scanned. Nonetheless, conventional TV remains the norm for the present. Even in the future, it will represent an educational historical study of how a significant problem was solved.

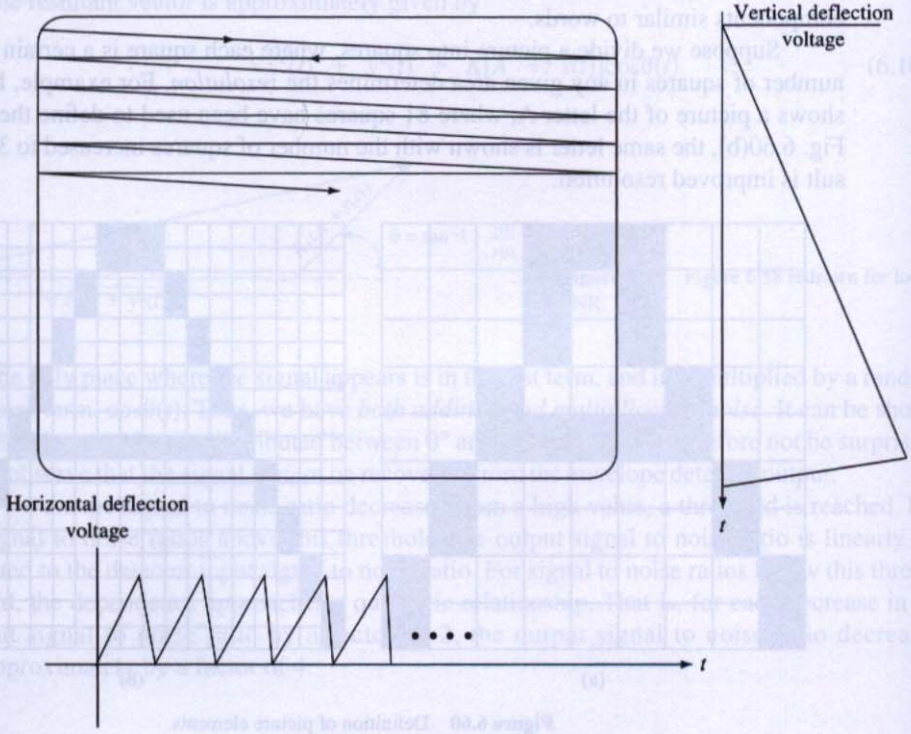


Figure 6.61 Generation of TV raster.

The screen has 495 horizontal lines in order to comply with the FCC regulation. After scanning the entire screen from top to bottom, the beam returns to the top of the screen, taking the equivalent time of an additional 30 lines to do so. We can therefore think of the picture as having 525 lines, 30 of which occur during the *vertical retrace* time.

We now assign timing to this process. The human eye requires a certain picture rate (the number of times the entire screen is traced each second) to avoid seeing flicker, as in early motion pictures. The minimum number is somewhere near 40 per second. United States TV uses 60 frames per second. (Color TV actually uses a number slightly less, approximately 59.94 Hz.) This matches the frequency of household current and is chosen so

as to minimize the effects of the video equivalent of 60-Hz hum. That is, if the 60-Hz power signal is not completely filtered out of the video, it causes a slight gradation of brightness over the height of the picture. If this gradation is stationary, the eye probably will not notice it. However, with a frequency mismatch, the gradation will exhibit a migration in the vertical direction (a rolling), and the likelihood of detecting it increases. We emphasize here that the numbers being presented are the U.S. standard (this applies to the United States, Canada, the Netherlands, Brazil, Colombia, Cuba, Japan, Mexico, Peru, Surinam, and Venezuela). Many of the standards in use in other parts of the world include 625 lines and a frame frequency of 50 Hz.

At 525 lines/frame and 60 frames/sec, the product of these is 31,500 lines/sec. The reciprocal of this is the time per line, 31.75 μsec/line. Of this 5.1 μsec are used for horizontal retrace, leaving 26.65 μsec for the visible part of each horizontal line. Dividing this by the 426 elements in a line yields the time per element of 0.0625 μsec/element, or 16 million elements/sec. The system therefore must be capable of transmitting 16 million independent shades (black, white, gray, and so on) per second. In the worst case, we may wish to display a perfect checkerboard design of alternating black and white squares. The system would then jump from the darkest to lightest shades and back again 16 million times a second.

If we now think of this light intensity information as a signal, we see from Fig. 6.62 that it has a fundamental frequency of 8 MHz. The square wave actually has a maximum frequency of infinity, but if we round it to a sinusoid, the human eye would still see the checkerboard design. Thus, no matter what modulation scheme we use to transmit this pattern, at least 8 MHz of bandwidth is required.

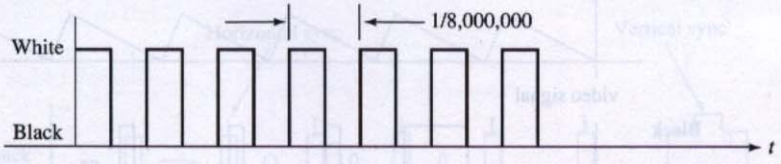


Figure 6.62 Intensity signal for checkerboard pattern.

Here the law steps in again, and the FCC mandated a maximum video signal bandwidth of 4.2 MHz. Alas, how can these seemingly contradictory specifications (60 frames/sec to avoid flicker, 211,000 picture elements for proper resolution, and a 4.2-MHz maximum bandwidth) be met?

Engineers had observed many decades of the development of motion pictures in which a similar predicament occurred. In standard motion pictures, only 24 different pictures are shown each second. But 24 flashes/sec on the screen would appear to flicker considerably. Contemporary motion picture projectors flash each image twice. (The film moves into position and the shutter opens and closes twice before the film moves again.) Thus, the frame rate is 48/sec, although the rate of presenting new pictures is 24/sec.

Television's founders decided to play a similar trick. They cut the signal frequency in half by cutting the number of lines per second in half, from 31,500 to 15,750 lines/sec. However, it was necessary to fill the entire screen each $\frac{1}{60}$ sec to avoid flicker. The technique for cutting the line frequency in half without changing the frame frequency is known as *interlaced scanning*. During the first $\frac{1}{60}$ sec, the odd-numbered lines are traced

(ending with $\frac{1}{2}$ line). The beam then returns to the top center of the screen to trace the even-numbered lines in the next $\frac{1}{6}$ sec. Thus, while it takes $\frac{1}{30}$ sec to send the frame consisting of all 211,000 picture elements, the screen is scanned twice (each scan is called a *field*) during this period. The eye fills in the missing rows and detects no flicker. There is some loss of resolution on fast-moving objects, but this was deemed appropriate for conventional TV.

Signal Design and Transmission

We now translate the foregoing information into an electrical signal format. If we plot light intensity as a function of time, a staircase function results. Figure 6.63 shows an example of the letter T in dark black, followed by a period in light gray. The associated signal is also shown; for simplicity, interlaced scanning has not been included, and the number of lines has been drastically reduced.

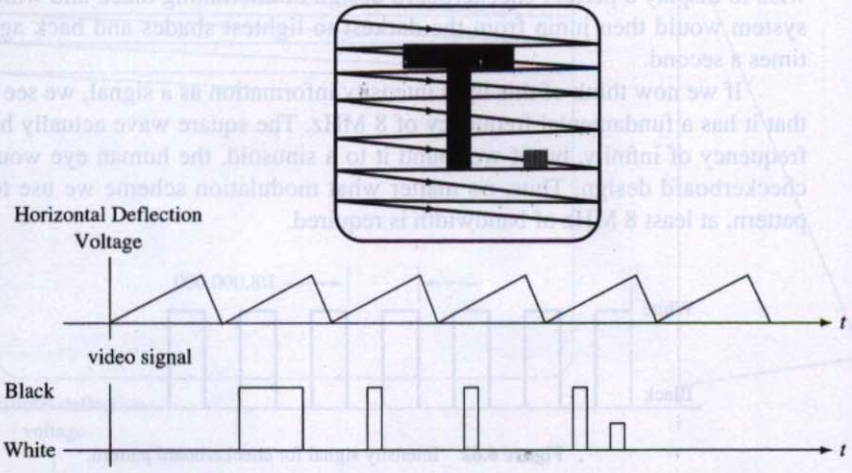


Figure 6.63 Video signal for particular message.

In broadcast TV, the information (video) signal would be a similar staircase function with minimum step width 0.125 μsec. In the actual video signal, the voltage corresponds to the light intensity, and the receiver uses this voltage to control the electron gun. The higher the voltage applied to a grid placed between the gun and screen, the fewer is the number of electrons that hit the screen and the darker is the spot. As an additional modification, the staircase function is smoothed to reduce the bandwidth. The eye cannot tell the difference between a smooth or rapid transition in the signal during 0.125 μsec. After all, this corresponds to only $\frac{1}{426}$ of the width of the TV screen.

There is an additional aspect to this electrical video signal known as *blanking*: While the beam on the cathode ray tube is retracing, it is desirable that the electron stream be turned off so that the retrace is not seen as a line on the screen.

Taking all this into account, the video signal corresponding to the picture shown in Fig. 6.63 is redrawn as Fig. 6.64.

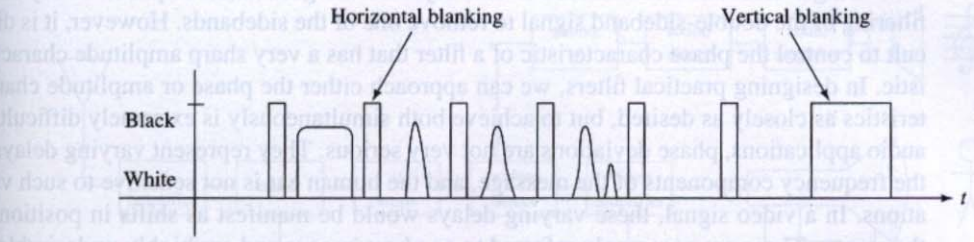


Figure 6.64 Video signal with addition of blanking.

Synchronization

The transmitter rapidly traces from left to right and from top to bottom many times each second. It sends a record of light intensity as a function of time. The receiver must be sure that it is placing the transmitted intensity element in the same location on the screen as intended in transmission. If the beam in the receiver does not start a scan at the same instant that the received waveform does, the picture will appear split at best and totally scrambled at worst. A method is thus required for synchronizing the two sweeping operations. This is done by means of synchronization pulses added to the video signal. The pulses are added during the blanking intervals, thereby not affecting what is seen on the screen. Figure 6.65 shows the signal of Fig. 6.64 modified with the addition of synchronization pulses. Two types of synchronizing pulses are shown in Fig. 6.65. The narrow pulses are horizontal synchronizing pulses, and the wide pulses are vertical synchronizing pulses.

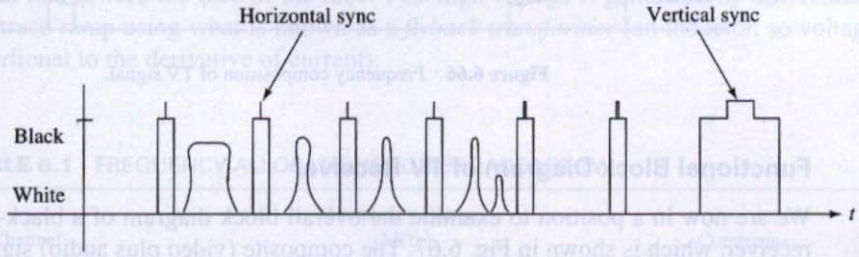


Figure 6.65 Video signal with addition of synchronization.

The receiver separates these pulses from the remaining signal through the use of a threshold circuit. The horizontal and vertical pulses are then distinguished by means of a single *RC* integrator circuit. The integral of the wider vertical sync pulses is larger than that of the narrower horizontal sync pulses. The separate pulses are then used to synchronize (trigger) the horizontal and vertical oscillators.

Modulation Techniques

The video signal has a maximum frequency of about 4 MHz. The FCC allocates 6 MHz of bandwidth to each television channel, and this space must contain both the video and audio sections of the transmitted signal. Obviously, the use of double-sideband AM must be rejected, since this would require over 8 MHz of bandwidth for each channel.

Single-sideband transmission is also rejected. Its generation requires a very sharp filtering of the double-sideband signal to remove one of the sidebands. However, it is difficult to control the phase characteristic of a filter that has a very sharp amplitude characteristic. In designing practical filters, we can approach either the phase or amplitude characteristics as closely as desired, but to achieve both simultaneously is extremely difficult. In audio applications, phase deviations are not very serious. They represent varying delays of the frequency components of the message, and the human ear is not sensitive to such variations. In a video signal, these varying delays would be manifest as shifts in position on the screen. These are commonly referred to as *ghost images* and are highly undesirable.

The video portion of the TV signal is sent using vestigial sideband. A carrier is added that is large enough to allow the use of an envelope detector for demodulation. The entire upper sideband and a portion of the lower sideband are sent.

Figure 6.66 shows the frequency composition of a TV signal. Note that the audio and video are frequency multiplexed, and their carriers are separated by 4.5 MHz. The audio is sent using FM, which is described in Chapter 7.

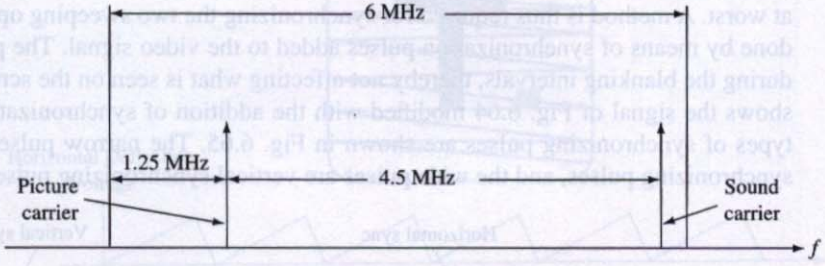


Figure 6.66 Frequency composition of TV signal.

Functional Block Diagram of TV Receiver

We are now in a position to examine the overall block diagram of a black-and-white TV receiver, which is shown in Fig. 6.67. The composite (video plus audio) signal is received by the antenna and is amplified by an RF amplifier. It then enters the tuner, which includes a mixer and an IF filter, just as in the case of the superheterodyne radio receiver. The frequency band allocated to commercial TV in the United States is shown in Table 6.1.

Most large cities with many active TV stations use channels 2, 4, 5, 7, 9, 11, and 13. Examination of the table of frequency allocations shows that this choice leaves a frequency separation between adjacent active stations, thus easing the requirements for design of the IF filters.

The IF frequency used for TV is 40 MHz. After IF amplification and filtering, the signal enters the video detector, which is simply an envelope detector. During this stage of processing, the sound signal is separated. Since the sound carrier is 4.5 MHz above the picture carrier, a filter (called a *sound trap*) is used to separate the sound signal from the video signal. The sound is sent by FM, so we defer discussion of the demodulating and processing of it until Chapter 7.

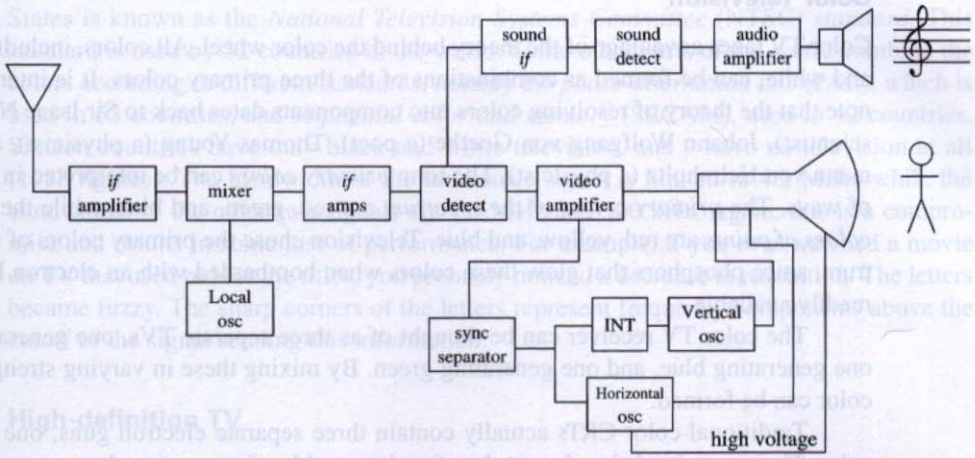


Figure 6.67 Block diagram of monochrome TV receiver.

The synchronizing pulses are separated from the video signal by means of a threshold circuit (clipper) known as the *sync separator*. The vertical sync pulses are then distinguished from the horizontal by an integrator. Both sets of pulses are used to trigger the corresponding sweep oscillators. As an added bonus, the retrace portion of the horizontal deflection voltage (with a very large slope due to a 5.1-μsec retrace time) is used to generate the very high voltage (over 20 kV) required on the CRT anode to pull the electrons in a straight line toward the face of the tube. This high voltage is generated by differentiating the retrace ramp using what is known as a *flyback transformer* (an inductor, so voltage is proportional to the derivative of current).

TABLE 6.1 FREQUENCY ALLOCATIONS FOR BROADCAST TV

Channel	Frequency Range (MHz)	Comments
2	54–60	
3	60–66	
4	66–72	
5	76–82	Note gap between 4 and 5
6	82–88	Between 6 and 7 is FM radio, aircraft, government, railroad, and police
7	174–180	
8	180–186	
9	186–192	
10	192–198	
11	198–204	
12	204–210	
13	210–216	
14–83	470–890	

Color Television

Color TV takes advantage of the theory behind the color wheel. All colors, including black and white, can be formed as combinations of the three primary colors. It is interesting to note that the theory of resolving colors into components dates back to Sir Isaac Newton (a scientist), Johann Wolfgang von Goethe (a poet), Thomas Young (a physicist), and Hermann von Helmholtz (a physicist). The term *primary colors* can be interpreted in a couple of ways. The *primary colors of the spectrum* are red, green, and blue, while the *primary colors of paints* are red, yellow, and blue. Television chose the primary colors of the spectrum, since phosphors that glow these colors when bombarded with an electron beam are readily available.

The color TV receiver can be thought of as three separate TVs, one generating red, one generating blue, and one generating green. By mixing these in varying strengths, any color can be formed.

Traditional color CRTs actually contain three separate electron guns, one for each color. The transmitted signal must therefore be capable of generating three separate video signals to control each of the guns.

But certainly, you are asking where the two additional signals can be squeezed. Monochrome transmission already uses all of the bandwidth allowed for TV. We need a system to transmit all three video signals without increasing the 6-MHz bandwidth, yet at the same time allowing compatible transmission. That is, a black-and-white TV receiver should receive the correct composite image, not just one of the colors.

The answer to this dilemma lies in a frequency analysis of the video signal. If we examine the Fourier transform of a black-and-white video signal, we find that it resembles a train of impulses. An actual signal for one frame contains 525 horizontal traces. Since most pictures contain some form of *vertical continuity* (that is, the content of one horizontal line closely resembles that of the next horizontal line), the video signal is almost periodic with a fundamental frequency of 15,750 Hz (the number of lines per second).

If the video signal were exactly periodic, its Fourier transform would be a train of impulses at multiples of the fundamental frequency. Since it is almost periodic, the Fourier transform consists of pulses (not impulses) centered around multiples of the line frequency. The more periodic the time signal, the sharper are the pulses in frequency.

Since the Fourier transform approaches zero between multiples of the line frequency, additional information can be placed in these spaces through the use of a form of frequency multiplexing. The additional information needed for color transmission modulates a carrier that is midway between two multiples of the line frequency. This is assured by using a carrier with a frequency that is an odd multiple of half of the line frequency. The figure used is 3.579545 MHz, which is the 455th harmonic of half of the line frequency. The two additional signals are combined and transmitted using quadrature AM.

To make the signal compatible with monochrome receivers, the three signals are *not* the three primary colors. Instead, the basic system transmits brightness (known as *luminance*), position in the color spectrum (*hue*), and how close the color is to a pure single frequency (*saturation*). The hue and saturation are combined with the luminance in a matrix operation that has the goal of making it easy to reproduce the three separate colors at the receiver and, at the same time, matching characteristics of our human color perception. The system does not provide perfect separation, and if the picture lacks vertical continuity

over a span of time, the colors will interact. The combination scheme used in the United States is known as the *National Television Systems Committee* (NTSC) *standard*. This standard is used by 32 countries in the world. Some other parts of the world combine the colors according to different standards, notably the *phase-alternation line* (PAL), which is used in 63 countries, and *sequential color and memory* (SECAM), used in 42 countries. Sixteen countries have only black-and-white television, and 9 have no television at all (1988 figures). The bandwidth of the luminance signal is limited to 4.2 MHz, while the bandwidths of the other two signals are 1.5 MHz and 500 kHz. Again, this is a compromise and results in some loss of performance. For example, if you ever watched a movie on TV that used red for the titles, you probably noticed a decrease in resolution. The letters became fuzzy. The sharp corners of the letters represent frequency components above the cutoff of the signal sending this information.

High-definition TV

At the time that standards were developed for television, few people dreamed of its evolution into a type of universal communication terminal. While these standards are acceptable for entertainment video, they are not sufficient for many evolving applications, such as videotext. For those, we require a high-resolution standard. *High-definition TV* (HDTV) is a term applied to a broad class of new systems. These systems have received worldwide attention. Indeed, the U.S. government put seed money into this consumer-related development in the hope of generating global competition and to be a catalyst in spurring the growth of new systems.

Of course, if we want to start from scratch, we can set the bandwidth of each channel to a number greater than 6 MHz, thereby achieving higher resolution. In fact, the Japan Broadcasting Corporation has done just that by assigning 10 MHz per channel and using compression techniques to achieve further improvement. The Japanese system permits 1,125 lines per frame, with 30 frames per second as 60 fields per second. (We will discuss compression techniques in Chapter 9.)

In the United States, the FCC has ruled that any new HDTV system must permit continuation of service to contemporary NTSC receivers. This significant constraint applies to terrestrial broadcasting (as opposed to videodisc, videotape, and cable television).

Developments in digital signal processing and high-speed RAMs have opened up interesting possibilities for increasing resolution while staying within the 6-MHz allocation per channel. For example, the number of lines could be increased if the frame rate were decreased. If the parameters for each pixel are stored, processing can be performed between frames. In the simplest example, the system could interpolate values between frames and create estimates of intermediate values. Indeed, more sophisticated compression algorithms can now be implemented, including variations of the powerful techniques currently applied to facsimile.

Other HDTV systems relax the 6-MHz constraint. For example, a VHF channel could be combined with a UHF channel, thus providing a total bandwidth of 12 MHz. Other systems involve transmission from direct-broadcast satellites. The one common thread among the various proposals is that the number of lines per frame is generally twice the current rate.

As of this writing, none of the competing systems has received anything near uniform support. The situation is such that the technical aspects of the problem are merged with social and economic issues. We can only hope that we do not end up with a multitude of incompatible systems.

PROBLEMS

6.2.1 Given an information signal $r(t)$, with

$$R(f) = A(f)e^{j\theta(f)}$$

(i.e., $R(f)$ is complex), find the Fourier transform of

$$r(t)\cos 2\pi f_c t$$

Also, find the Fourier transform of

$$r(t)\cos\left(\frac{2\pi f_c t + \pi}{4}\right)$$

6.3.1 The signal shown in Fig. P6.3.1 amplitude modulates a carrier of frequency 10^6 Hz.

(a) If the modulation is DSBSC, sketch the modulated waveform.

(b) The modulated wave of part (a) forms the input to an envelope detector. Sketch the output of the detector.

(c) A carrier term is now added to the DSBSC waveform. What is the minimum amplitude of the carrier such that envelope detection can be used?

(d) For the modulated signal of part (c), sketch the output of an envelope detector.

(e) Draw a block diagram of a synchronous detector that could be used to recover $s(t)$ from the modulated waveform of part (a).

(f) Sketch the output of the synchronous demodulator if the waveform of part (c) forms the input.

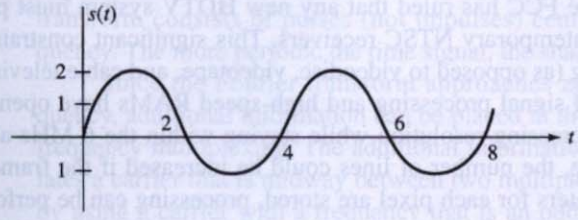


Fig. P6.3.1

6.4.1 You are given the voltage signals $s(t)$ and $\cos 2\pi f_c t$, and you wish to produce the AM wave. Discuss two practical methods of generating this AM waveform.

6.4.2 A system is as shown in Fig. P6.4.2. Note that the system resembles a gated modulator, except that the gating function goes between $+1$ and -1 instead of between $+1$ and 0, and the bandpass filter has been replaced with a lowpass filter.

Can this system still produce an AM waveform? If your answer is yes, find the minimum and maximum values of f_{LPF} in order for the system to act as a modulator. If your answer is no, show all work that made you arrive at this conclusion.

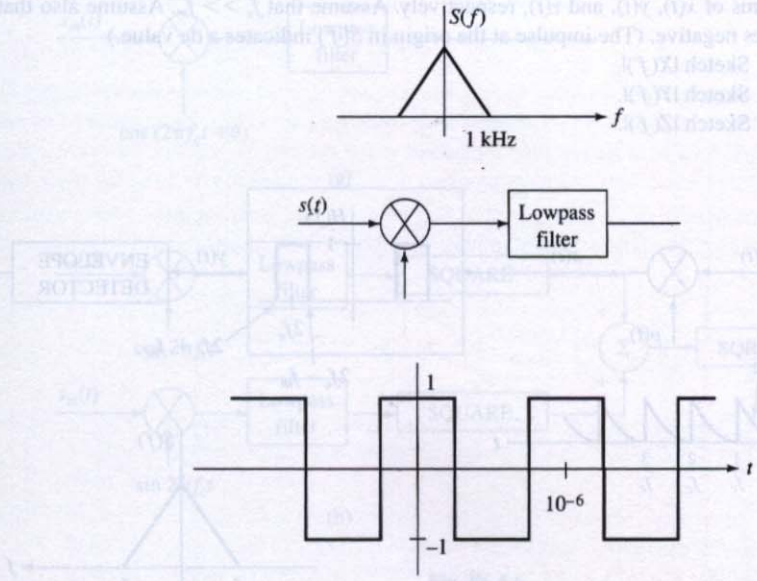


Fig. P6.4.2

6.5.1 The signal

$$s(t) = \frac{2\sin 2\pi t}{t}$$

modulates a carrier of frequency 100 Hz. A signal of the form

$$n(t) = \frac{\sin 199\,\pi t}{t}$$

adds to the AM waveform, and the sum forms the input to a synchronous demodulator. Find the output of the demodulator.

6.5.2 The waveform $v_{in}(t)$ shown in Fig. P6.5.2 forms the input to an envelope detector. Sketch the output waveform.

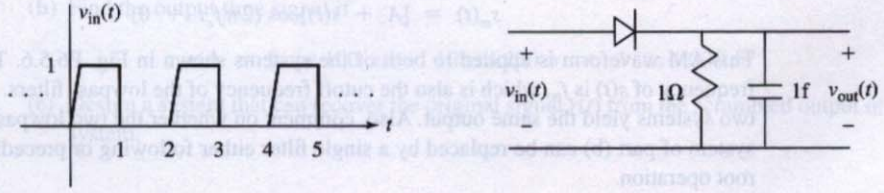


Fig. P6.5.2

6.5.3 You are given the system shown in Fig. P6.5.3. An AMTC waveform forms the input. $p(t)$ is the periodic function, and $S(f)$ is as sketched. $X(f)$, $Y(f)$, and $Z(f)$ are the Fourier trans-

forms of $x(t)$, $y(t)$, and $z(t)$, respectively. Assume that $f_c >> f_m$. Assume also that $s(t)$ never goes negative. (The impulse at the origin in $S(f)$ indicates a dc value.)

(a) Sketch $|X(f)|$.

(b) Sketch $|Y(f)|$.

(c) Sketch $|Z(f)|$.

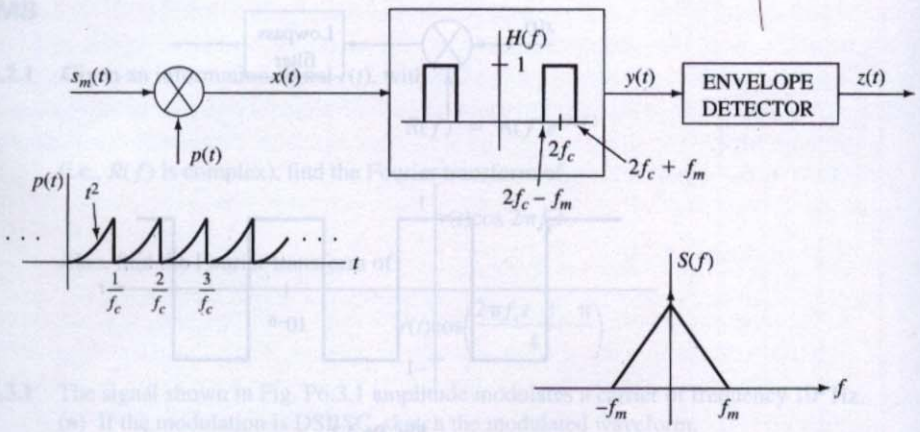


Fig. P6.5.3

6.5.4 The input to an envelope detector is

$$r(t)\cos 2\pi f_c t$$

where $r(t)$ is always greater than zero.

(a) What is the output of the envelope detector?

(b) What is the average power of the input in terms of the average power of $r(t)$?

(c) What is the average power of the output?

(d) Discuss any apparent discrepancies.

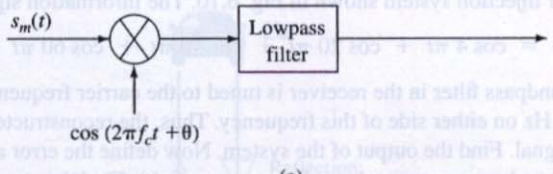
6.5.5 Replace the local carrier in a synchronous demodulator with a square wave at a fundamental frequency of f_c. Will the system still operate as a demodulator? Will the same be true if periodic signals other than the square wave are substituted for the oscillator?

6.5.6 An AMTC signal $s_m(t)$ is given by

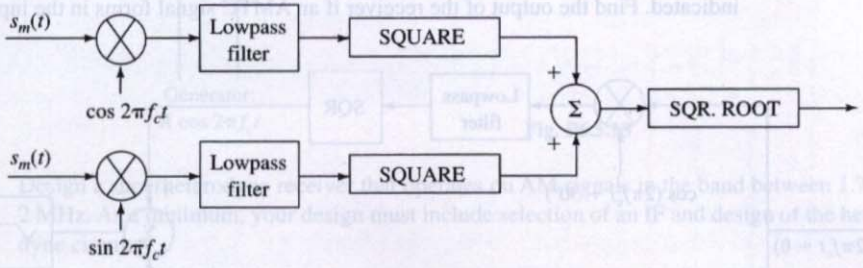
$$s_m(t) = [A + s(t)]\cos(2\pi f_c t + \theta)$$

This AM waveform is applied to both of the systems shown in Fig. P6.5.6. The maximum frequency of $s(t)$ is f_m, which is also the cutoff frequency of the lowpass filters. Show that the two systems yield the same output. Also, comment on whether the two lowpass filters in the system of part (b) can be replaced by a single filter either following or preceding the square root operation.

6.5.7 A synchronous demodulator is used to detect an amplitude-modulated suppressed carrier double-sideband waveform. In designing the detector, the frequency is matched perfectly, but the phase differs from that of the received carrier by $\Delta\theta$, as shown in Fig. P6.5.7. The phase difference is random and Gaussian distributed with a mean of zero and variance of σ^2. When the phases are matched, the output is $s(t)/2$. What is the maximum variance of the phase error such that the output amplitude is at least 50% of this optimum value 99% of the time?



(a)



(b)

Fig. P6.5.6

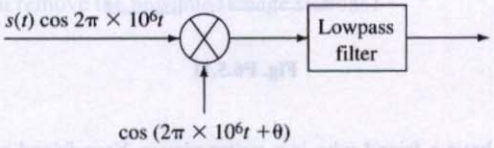


Fig. P6.5.7

6.5.8 Show that the rectifier detector of Fig. 6.32 demodulates transmitted carrier AM when the rectifier is half wave.

6.5.9 The system illustrated in Fig. P6.5.9 performs a simple scrambling operation: reversing frequencies. (That is, dc switches to the highest frequency, while the highest frequency switches to dc; frequencies near f_m flip to a location near zero.)

(a) Sketch the Fourier transform of the output signal $Y(f)$.

(b) Find the output time signal if

$$r(t) = 5\cos 100\pi t + 10\cos 200\pi t + 3\cos 1{,}000\pi t$$

(c) Design a system that can recover the original signal $r(t)$ from the scrambled output of the system.

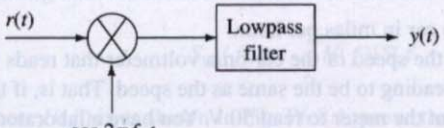


Fig. P6.5.9

6.5.10 Consider the carrier injection system shown in Fig. 6.10. The information signal is

$$r(t) = \cos 4\pi t + \cos 20\pi t + \cos 40\pi t + \cos 60\pi t$$

Suppose that the bandpass filter in the receiver is tuned to the carrier frequency with a bandwidth that passes 3 Hz on either side of this frequency. Thus, the reconstructed carrier is perturbed by a 2-Hz signal. Find the output of the system. Now define the error as the difference between this output and an appropriately scaled version of $r(t)$. Find the percentage of error.

6.5.11 A quadrature receiver is shown in Fig. P6.5.11. The local oscillators have shifted by 30°, as indicated. Find the output of the receiver if an AMTC signal forms in the input as shown.

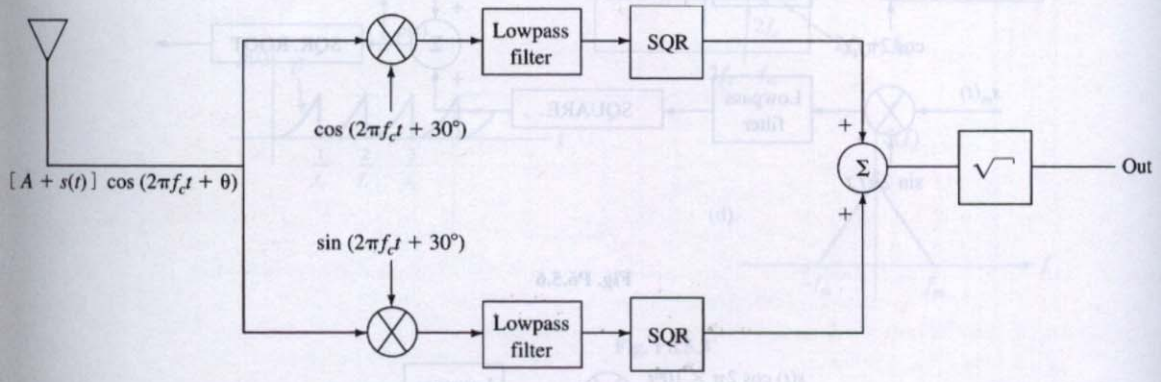


Fig. P6.5.11

6.5.12 You have a friend who is a guitar player. Your friend asks you to design a system that will help in the tuning of a guitar. Specifically, you are asked to design a system that accepts two inputs,

$$\cos(2\pi f_0 t); \quad \cos[2\pi(f_0 + \Delta f)t + \theta]$$

and produces an output which is a dc signal that is proportional to the frequency difference Δf. The phase angle θ is unknown. Draw a block diagram of your system.

6.5.13 You wish to design a *Doppler radar system*. A sinusoidal generator continuously generates the signal,

$$s(t) = A\cos 2\pi f_c t$$

This signal is transmitted to a speeding car, and the reflected signal is of the form

$$r(t) = B\cos[2\pi(f_c + \Delta f)t]$$

The situation is shown in Fig. P6.5.13. The frequency difference is

$$\Delta f = 10s$$

where s is the speed of the car in miles per hour.

You wish to display the speed of the car on a voltmeter that reads from 0 to 100 volts, and you wish the voltage reading to be the same as the speed. That is, if the car is traveling at 50 miles per hour, you want the meter to read 50 V. You have a laboratory full of equipment, including filters, multipliers, and any other type of device you need. Design the system.

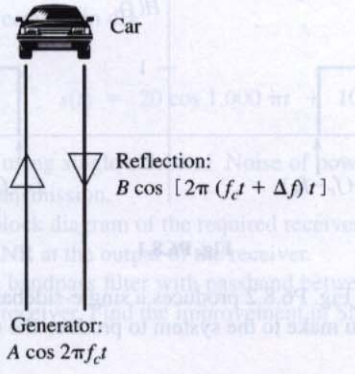


Fig. P6.5.13

6.6.1 Design a superheterodyne receiver that operates on AM signals in the band between 1.7 and 2 MHz. As a minimum, your design must include selection of an IF and design of the heterodyne circuitry.

6.6.2 Figure P6.6.2 shows a *dual conversion receiver*. Assume that the first IF frequency is 30 MHz and the second is 10 MHz. Assume further that the receiver is designed to demodulate a band of channels between 135 and 136 MHz, each of which is 100 kHz in bandwidth.

(a) Suggest the range of frequencies for the local oscillators.

(b) Determine all possible image station frequencies.

(c) How would you remove the unwanted image stations?

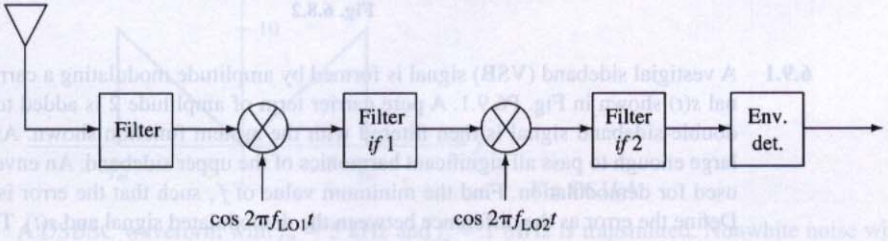


Fig. P6.6.2

6.6.3 Twenty-five radio stations are broadcasting in the band between 3 MHz and 3.5 MHz. You wish to modify an AM broadcast receiver to receive the broadcasts. Each audio signal has a maximum frequency $f_m = 10$ kHz. Describe, in detail, the changes you would have to make to the standard broadcast superheterodyne receiver in order to receive the broadcast.

6.7.1 Determine the envelope of the waveform

$$s(t) = \cos 10\,\pi t + 17 \cos 30\,\pi t \cos 1000\,\pi t$$

6.8.1 Starting with the transform of a lower sideband single-sideband wave

$$S_{lsb}(f) = \frac{1}{2} H(f)[S(f - f_c) + S(f + f_c)]$$

where $H(f)$ is as shown in Fig. P6.8.1, prove that synchronous demodulation can be used to recover $s(t)$ from $s_{lsb}(t)$.

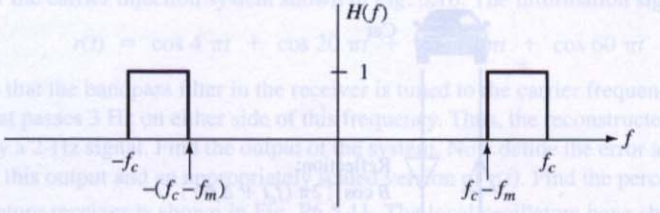


Fig. P6.8.1

6.8.2 Show that the system of Fig. P6.8.2 produces a single-sideband (lower sideband) waveform. What changes would you make to the system to produce the upper sideband waveform?

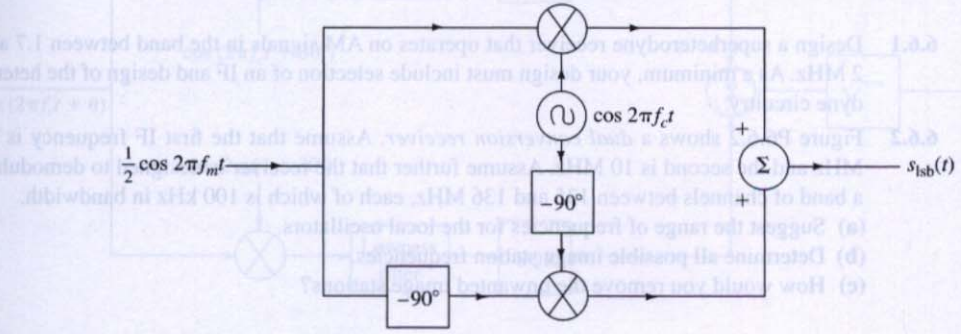


Fig. 6.8.2

6.9.1 A vestigial sideband (VSB) signal is formed by amplitude modulating a carrier with the signal $s(t)$ shown in Fig. P6.9.1. A pure carrier term of amplitude 2 is added to the result. The double-sideband signal is then filtered with the system function shown. Assume that f_m is large enough to pass all significant harmonics of the upper sideband. An envelope detector is used for demodulation. Find the minimum value of f_v such that the error is less than 10%. Define the error as the difference between the demodulated signal and $s(t)$. The percent error is the ratio of error power to signal power.

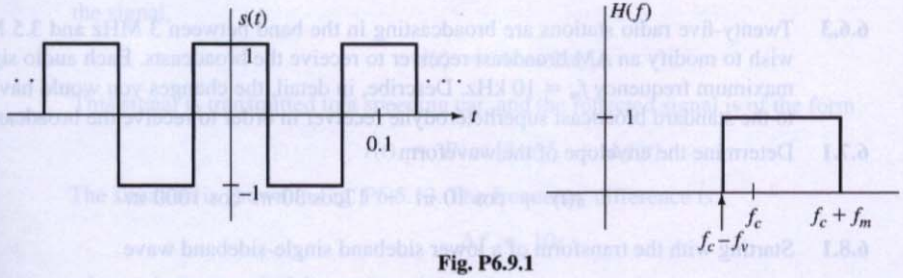


Fig. P6.9.1

6.11.1 An information signal $s(t) = 5\cos 1{,}000\pi t$ is transmitted using single-sideband suppressed carrier and is demodulated using a synchronous demodulator. Noise with power spectral

density $G_n(f) = N_0/2 = 10^{-4}$ adds to the signal during transmission. Find the SNR at the output of the receiver, in dB.

6.11.2 A signal

$$s(t) = 20 \cos 1{,}000\,\pi t + 10 \cos 2{,}000\,\pi t$$

is transmitted using single sideband. Noise of power spectral density $G_n(f) = N_0/2 = 10^{-3}$ adds during transmission.

(a) Sketch a block diagram of the required receiver.
(b) Find the SNR at the output of the receiver.
(c) Suppose a bandpass filter with passband between 400 and 1,100 Hz is added to the output of the receiver. Find the improvement in SNR of this filter.

6.11.3 A signal

$$s(t) = 4 \sin (200\,\pi t + 10°)$$

is transmitted using double-sideband transmitted carrier, double-sideband suppressed carrier, and single sideband. Noise of power spectral density $G_n(f) = N_0/2 = 10^{-2}$ adds during transmission. Find the SNR at the output of the appropriate receiver for each case. (Assume that an envelope detector is used for double-sideband transmitted carrier.)

6.11.4 A signal $s(t)$ is transmitted using single-sideband AM. The power spectral density of $s(t)$ is as shown in Fig. P6.11.4. White noise of spectral density $N_0/2$ adds during transmission. Find the SNR at the output of a synchronous demodulator.

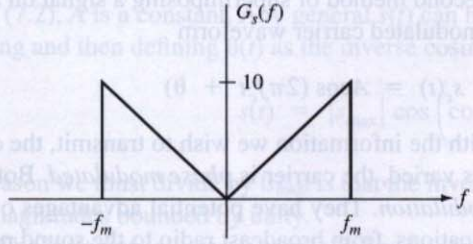


Fig. P6.11.4

6.11.5 A DSBSC waveform with $f_m = 5$ kHz and $f_c = 1$ MHz is transmitted. Nonwhite noise with power spectral density as shown in Fig. P6.11.5 adds to the signal prior to detection with a synchronous demodulator. Find the SNR at the output of the detector, assuming that the power of the signal, $s(t)$, is 1 watt.

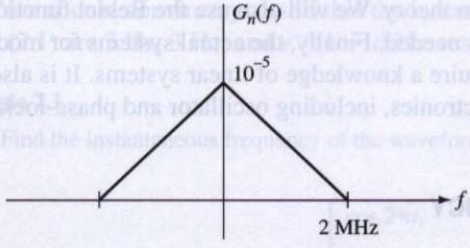


Fig. P6.11.5

6.12.1 Discuss the trade-off decision required to increase the vertical resolution of commercial TV by 10%.