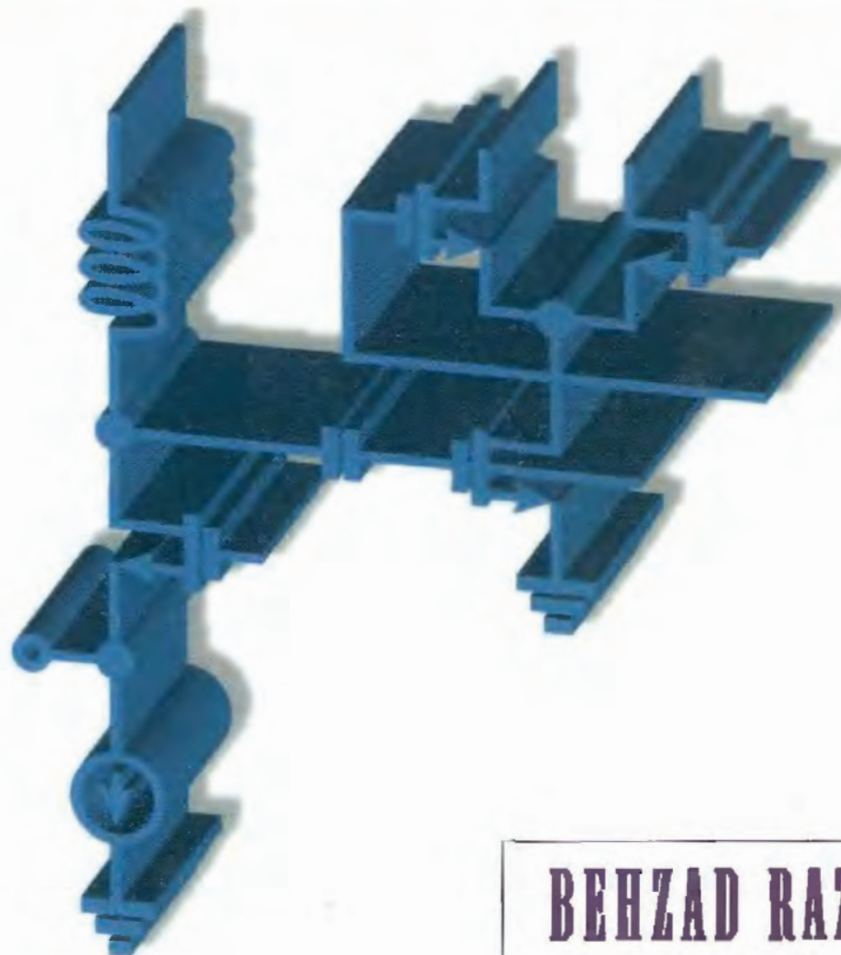


RF

MICROELECTRONICS



BEHZAD RAZAVI

PRENTICE HALL COMMUNICATIONS ENGINEERING AND EMERGING TECHNOLOGIES SERIES
Theodore S. Rappaport, Series Editor

EX-1030
U.S. Patent 8,401,875

RF MICROELECTRONICS

ISBN 0-13-887571-5



90000

9 780138 875718

PRENTICE HALL COMMUNICATIONS ENGINEERING
AND EMERGING TECHNOLOGIES SERIES

Theodore S. Rappaport, Series Editor

RAPPAPORT *Wireless Communications: Principles and Practice*

RAZAVI *RF Microelectronics*

Forthcoming:

LIBERTI & RAPPAPORT *CDMA and Adaptive Antennas for Wireless Systems*

TRANter, KURT, KOSBAR, & RAPPAPORT *Simulation of Modern Communications
Systems with Wireless Applications*

GARG & WILKES *Global System Mobile Communication*

RF MICROELECTRONICS

Behzad Razavi
University of California, Los Angeles

To join a Prentice Hall PTR
Internet mailing list, point to
http://www.prenhall.com/mail_lists



PRENTICE HALL PTR
Upper Saddle River, NJ 07458

Library of Congress Cataloging-in-Publication Data

RAZAVI, BEHZAD.

RF microelectronics / Behzad Razavi.

p. cm. – (Prentice Hall communications engineering and emerging technologies series ; 2)

Includes bibliographical references and index.

ISBN 0-13-887571-5

1. Radio circuits—Design and construction 2. Integrated circuits—Design and construction I. Title. II. Series.

TK6560.R39 1998

97-34526

621.384'12—dc21

CIP

Acquisitions Editor: *Russ Hall*

Editorial Assistant: *Maureen Diana*

Editorial/Production Supervision: *Kerry Reardon*

Manufacturing Manager: *Alexis R. Heydt*

Cover Design: *Alamini Design*

Cover Design Director: *Jerry Votta*

Composition: *PreTeX, Inc.*



© 1998 Prentice Hall PTR

Prentice-Hall, Inc.

Upper Saddle River NJ 07458

Prentice Hall books are widely used by corporations and government agencies for training, marketing, and resale. The publisher offers discounts on this book when ordered in bulk quantities.

For more information contact:

Corporate Sales Department

Phone: 800-382-3419 Fax: 201-236-7141

E-Mail: corpsales@prenhall.com

or write: Prentice Hall PTR

One Lake Street

Upper Saddle River, NJ 07458

All rights reserved. No part of this book may be reproduced, in any form or by any means, without permission in writing from the publisher.

Printed in the United States of America

17 18 19

ISBN 0-13-887571-5

Prentice-Hall International (UK) Limited, *London*

Prentice-Hall of Australia Pty. Limited, *Sydney*

Prentice-Hall of Canada, Inc., *Toronto*

Prentice-Hall Hispanoamericana, S. A., *Mexico*

Prentice-Hall of India Private Limited, *New Delhi*

Prentice-Hall of Japan, Inc., *Tokyo*

Prentice-Hall Asia Pte. Ltd., *Singapore*

Editora Prentice-Hall do Brasil, Ltda., *Rio de Janeiro*

To the memory
of my parents

Contents

Preface	xiii
----------------	-------------

► 1 Introduction to RF and Wireless Technology	1
--	---

1.1 Complexity Comparison	2
1.2 Design Bottleneck	4
1.3 Applications	6
1.4 Analog and Digital Systems	7
1.5 Choice of Technology	9
<i>References</i>	<i>10</i>

► 2 Basic Concepts in RF Design	11
---------------------------------	----

2.1 Nonlinearity and Time Variance	11
2.1.1 Effects of Nonlinearity	14
2.1.2 Cascaded Nonlinear Stages	22
2.2 Intersymbol Interference	25
2.3 Random Processes and Noise	28
2.3.1 Random Processes	29
2.3.2 Noise	37
2.4 Sensitivity and Dynamic Range	48
2.5 Passive Impedance Transformation	50
<i>References</i>	<i>53</i>

► 3 Modulation and Detection	54
------------------------------	----

3.1 General Considerations	54
3.2 Analog Modulation	57
3.2.1 Amplitude Modulation	57
3.2.2 Phase and Frequency Modulation	58

3.3	Digital Modulation	63
3.3.1	Basic Concepts	64
3.3.2	Binary Modulation	74
3.3.3	Quadrature Modulation	81
3.4	Power Efficiency of Modulation Schemes	90
3.4.1	Constant- and Variable-Envelope Signals	90
3.4.2	Spectral Regrowth	91
3.5	Noncoherent Detection	93
	References	96

► 4 Multiple Access Techniques and Wireless Standards 98

4.1	Mobile RF Communications	98
4.2	Multiple Access Techniques	103
4.2.1	Time- and Frequency-Division Duplexing	103
4.2.2	Frequency-Division Multiple Access	105
4.2.3	Time-Division Multiple Access	105
4.2.4	Code-Division Multiple Access	107
4.3	Wireless Standards	110
4.3.1	Advanced Mobile Phone Service	111
4.3.2	North American Digital Standard	112
4.3.3	Global System for Mobile Communication	113
4.3.4	Qualcomm CDMA	114
4.3.5	Digital European Cordless Telephone	116
	References	117

► 5 Transceiver Architectures 118

5.1	General Considerations	118
5.2	Receiver Architectures	122
5.2.1	Heterodyne Receivers	122
5.2.2	Homodyne Receivers	129
5.2.3	Image-Reject Receivers	138
5.2.4	Digital-IF Receivers	146
5.2.5	Subsampling Receivers	147
5.3	Transmitter Architectures	149
5.3.1	Direct-Conversion Transmitters	152
5.3.2	Two-Step Transmitters	154
5.4	Transceiver Performance Tests	155

5.5	Case Studies	157
5.5.1	Motorola's FM Receiver	157
5.5.2	Philips' Pager Receiver	158
5.5.3	Philips' DECT Transceiver	159
5.5.4	Lucent Technologies' GSM Transceiver	161
5.5.5	Philips' GSM Transceiver	162
	References	163

►	6	Low-Noise Amplifiers and Mixers	166
---	---	---------------------------------	-----

6.1	Low-Noise Amplifiers	166
6.1.1	General Considerations	166
6.1.2	Input Matching	170
6.1.3	Bipolar LNAs	173
6.1.4	CMOS LNAs	178
6.2	Downconversion Mixers	180
6.2.1	General Considerations	180
6.2.2	Bipolar Mixers	188
6.2.3	CMOS Mixers	192
6.2.4	Noise in Mixers	194
6.3	Cascaded Stages Revisited	200
	References	204

►	7	Oscillators	206
---	---	-------------	-----

7.1	General Considerations	206
7.2	Basic LC Oscillator Topologies	209
7.3	Voltage-Controlled Oscillators	212
7.4	Phase Noise	214
7.4.1	Effect of Phase Noise in RF Communications	215
7.4.2	Q of an Oscillator	217
7.4.3	Phase Noise Mechanisms	220
7.4.4	Noise-Power Trade-off	224
7.4.5	Effect of Frequency Division and Multiplication on Phase Noise	224
7.4.6	Oscillator Pulling and Pushing	225
7.5	Bipolar and CMOS LC Oscillators	227
7.5.1	Negative- G_m Oscillators	227
7.5.2	Interpolative Oscillators	231

- 7.6 Monolithic Inductors 233
- 7.7 Resonatorless VCOs 235
- 7.8 Quadrature Signal Generation 236
 - 7.8.1 RC-CR Network 236
 - 7.8.2 Havens' Technique 239
 - 7.8.3 Frequency Division 243
- 7.9 Single-Sideband Generation 243
- References 244**

► **8 Frequency Synthesizers** 247

- 8.1 General Considerations 247
- 8.2 Phase-Locked Loops 249
 - 8.2.1 Basic Concepts 249
 - 8.2.2 Basic PLL 252
 - 8.2.3 Charge-Pump PLLs 258
 - 8.2.4 Type I and Type II PLLs 265
 - 8.2.5 Noise in PLLs 266
 - 8.2.6 Phase Noise at Input 266
 - 8.2.7 Phase Noise of VCO 267
 - 8.2.8 Frequency Multiplication 269
- 8.3 RF Synthesizer Architectures 269
 - 8.3.1 Integer- N Architecture 270
 - 8.3.2 Fractional- N Architecture 277
 - 8.3.3 Dual-Loop Architectures 284
 - 8.3.4 Direct Digital Synthesis 285
- 8.4 Frequency Dividers 290
 - 8.4.1 Divide-by-Two Circuits 290
 - 8.4.2 Dual-Modulus Dividers 293
- References 296**

► **9 Power Amplifiers** 298

- 9.1 General Considerations 298
 - 9.1.1 Linear and Nonlinear PAs 301
- 9.2 Classification of Power Amplifiers 301
 - 9.2.1 Class A and B PAs 302
 - 9.2.2 Class C PAs 305

9.3	High-Efficiency Power Amplifiers	306
9.4	Large-Signal Impedance Matching	310
9.5	Linearization Techniques	313
9.5.1	Feedforward	314
9.5.2	Feedback	315
9.5.3	Envelope Elimination and Restoration	317
9.5.4	LINC	318
9.6	Design Examples	320
	References	323

Preface

The annual worldwide sales of cellular phones has exceeded \$2.5B. With 4.5 million customers, home satellite networks comprise a \$2.5B industry. The global positioning system is expected to become a \$5B market by the year 2000. In Europe, the sales of equipment and services for mobile communications will reach \$30B by 1998. The statistics are overwhelming.

The radio frequency (RF) and wireless market has suddenly expanded to unimaginable dimensions. Devices such as pagers, cellular and cordless phones, cable modems, and RF identification tags are rapidly penetrating all aspects of our lives, evolving from luxury items to indispensable tools. Semiconductor and system companies, small and large, analog and digital, have seen the statistics and are striving to capture their own market share by introducing various RF products.

RF design is unique in that it draws upon many disciplines unrelated to integrated circuits (ICs). The RF knowledge base has grown for almost a century, creating a seemingly endless body of literature for the novice.

This book deals with the analysis and design of RF integrated circuits and systems. Providing a systematic treatment of RF electronics in a tutorial language, the book begins with the necessary background knowledge from microwave and communication theory and leads the reader to the design of RF transceivers and circuits. The text emphasizes both architecture and circuit level issues with respect to monolithic implementation in VLSI technologies. The primary focus is on bipolar and CMOS design, but most of the concepts can be applied to other technologies as well. The reader is assumed to have a basic understanding of analog IC design and the theory of signals and systems.

The book consists of nine chapters. Chapter 1 gives a general introduction, posing questions and providing motivation for subsequent chapters. Chapter 2 describes basic concepts in RF and microwave design, emphasizing the effects of nonlinearity and noise.

Chapters 3 and 4 take the reader to the communication system level, giving an overview of modulation, detection, multiple access techniques, and wireless standards. While initially appearing to be unnecessary, this material is in fact essential to the concurrent design of RF circuits and systems.

Chapter 5 deals with transceiver architectures, presenting various receiver and transmitter topologies along with their merits and drawbacks. This chapter

also includes a number of case studies that exemplify the approaches taken in actual RF products.

Chapters 6 through 9 address the design of RF building blocks: low-noise amplifiers and mixers, oscillators, frequency synthesizers, and power amplifiers, with particular attention to minimizing the number of off-chip components. An important goal of these chapters is to demonstrate how the system requirements define the parameters of the circuits and how the performance of each circuit impacts that of the overall transceiver.

I have taught approximately 80% of the material in this book in a 4-unit graduate course at UCLA. Chapters 3, 4, 8, and 9 had to be shortened in a ten-week quarter, but in a semester system they can be covered more thoroughly.

Much of my RF design knowledge comes from interactions with colleagues. Helen Kim, Ting-Ping Liu, and Dan Avidor of Bell Laboratories, and David Su and Andrew Gzegorek of Hewlett-Packard Laboratories have contributed to the material in this book in many ways. The text was also reviewed by a number of experts: Stefan Heinen (Siemens), Bart Jansen (Hewlett-Packard), Ting-Ping Liu (Bell Labs), John Long (University of Toronto), Tadao Nakagawa (NTT), Gitty Nasserbakht (Texas Instruments), Ted Rappaport (Virginia Tech), Tirdad Sowlati (Gennum), Trudy Stetzler (Bell Labs), David Su (Hewlett-Packard), and Rick Wesel (UCLA). In addition, a number of UCLA students, including Farbod Behbahani, Hooman Darabi, John Leete, and Jacob Rael, “test drove” various chapters and provided useful feedback. I am indebted to all of the above for their kind assistance.

I would also like to thank the staff at Prentice Hall, particularly Russ Hall, Maureen Diana, and Kerry Reardon for their support.

Behzad Razavi
July 1997

INTRODUCTION TO RF AND WIRELESS TECHNOLOGY

"This phone uses GFSK modulation in DECT-TDMA/TDD with zero-IF I/Q detection," said the engineer. "How can we modify it to work with DCS1800 as well?" asked the manager. "We would need to add a duplexer, at least one more LO with SSB mixing, and probably two SAW filters," replied the engineer . . .

Telephones have gotten much more complicated than they used to be. So have RF circuits. The nonspecialist who uses a cellular phone (cellphone) to call home probably does not know that hundreds of scientists and engineers have worked for almost a century to make wireless technology affordable. Nor does he know that there is more computing power in the phone than in some of the early personal computers.

Wireless technology came to existence when, in 1901, Guglielmo Marconi successfully transmitted radio signals across the Atlantic Ocean. The consequences and prospects of this demonstration were simply overwhelming; the possibility of replacing telegraph and telephone communications with wave transmission through the "ether" portrayed an exciting future. However, while two-way wireless communication did materialize in military applications, wireless transmission in daily life remained limited to one-way radio and television broadcasting by large, expensive stations. Ordinary two-way phone conversations would still go over wires for many decades.

The invention of the transistor, the development of Shannon's information theory, and the conception of the cellular system—all at Bell Laboratories—paved the way for affordable mobile communications, as originally implemented in car phones and eventually realized in portable cellular phones.

But why the sudden surge in wireless electronics? Market surveys show that in the United States more than 28,000 people join the cellular phone system *every day*, motivating competitive manufacturers to provide phone sets with increasingly higher performance and lower cost. In fact, the present goal

is to reduce both the power consumption and price of cellphones by 30% every year—although it is not clear for how long this rate can be sustained. A more glorious prospect, however, lies in the power of two-way wireless communication when it is introduced in other facets of our lives: home phones, computers, facsimile, and television. While an immediate objective of the wireless industry is to combine cordless and cellular phones so as to allow seamless communications virtually everywhere, the long-term plan is to produce an “omnipotent” wireless terminal that can handle voice, data, and video as well as provide computing power. Other luxury items such as the global positioning system (GPS) are also likely to become available through this terminal sometime in the future. Personal communication services (PCS) are almost here.

1.1 COMPLEXITY COMPARISON

To gain an early perspective on modern RF design, first consider the system depicted in Fig. 1.1, a simple frequency modulation (FM) transceiver. In Fig. 1.1(a), Q_1 operates as both an oscillator and a frequency modulator; that

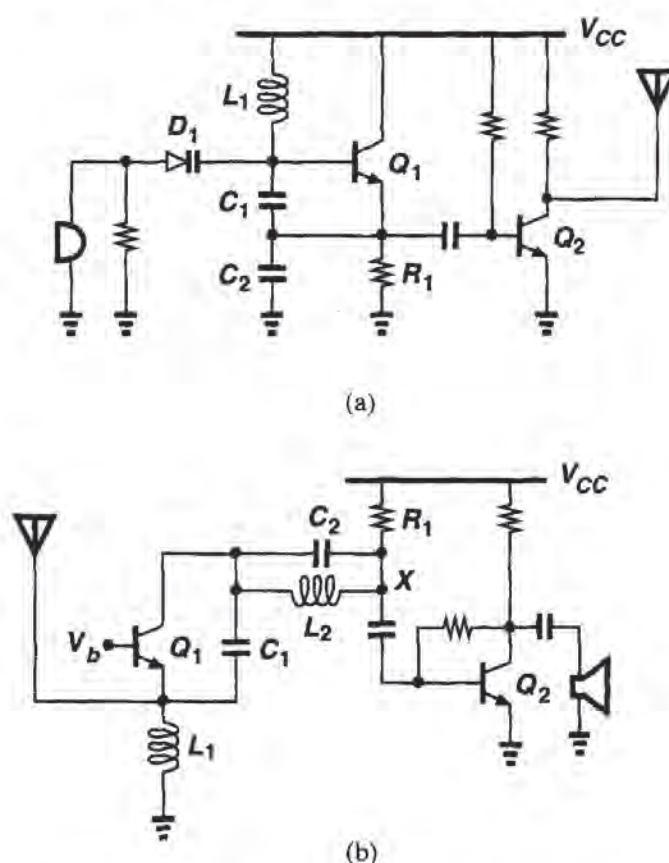


Figure 1.1 (a) FM transmitter, (b) FM receiver.

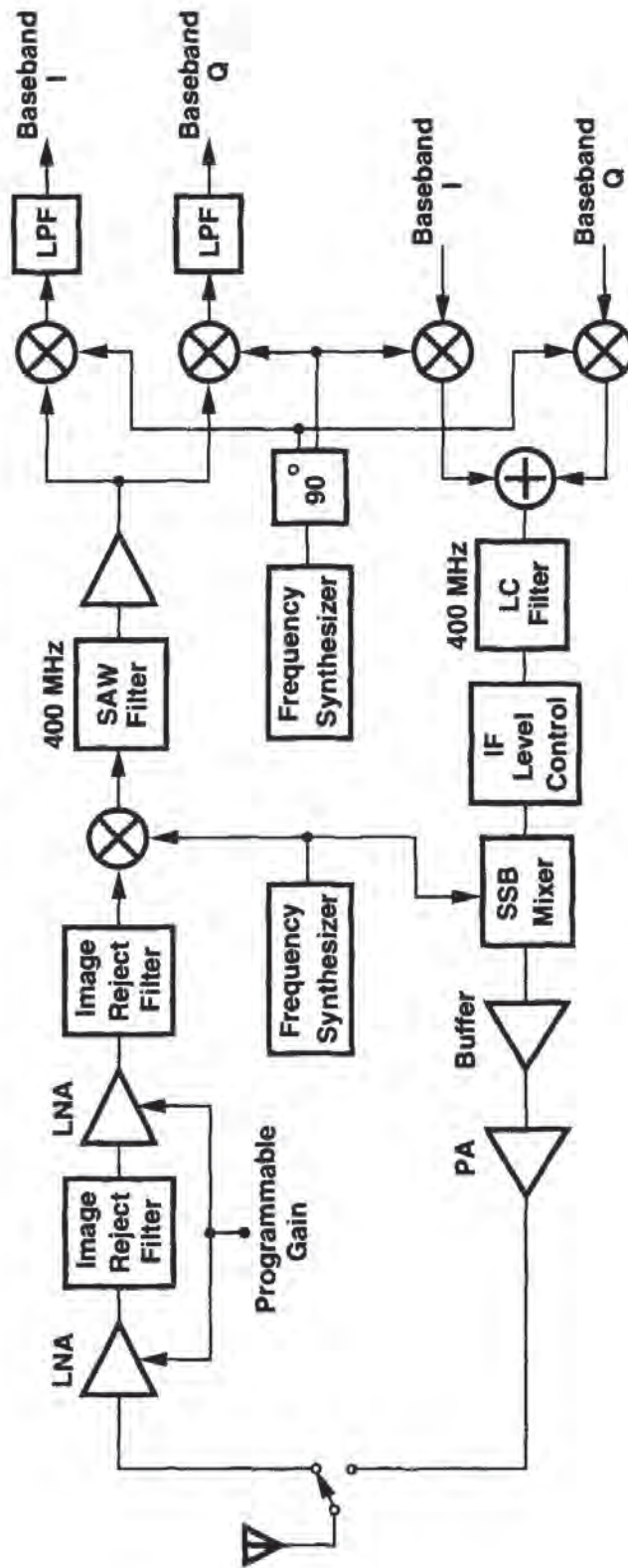


Figure 1.2 RF section of a cellphone [1]

is, the audio signal produced by the microphone varies the bias voltage across the varactor diode D_1 , thereby modulating the frequency of oscillation. In the receive path, Q_1 serves as both an oscillator and a demodulator, generating the audio signal at node X . The result is then amplified and applied to the speaker.

Now consider the circuit of Fig. 1.2, the RF section of a cellphone [1]. This circuit is orders of magnitude more complex than the FM circuits of Fig. 1.1, and we will postpone its analysis until Chapter 5. Why have RF designers gone from the circuit of Fig. 1.1 to that of Fig. 1.2? What is the thought process behind this evolution? Is all of this complexity really necessary? These questions will be answered in Chapters 2 through 5 as we develop our understanding of RF communication systems and architecture and circuit design issues.

1.2 DESIGN BOTTLENECK

Today's pocket phones contain more than one million transistors, with only a small fraction operating in the RF range and the rest performing low-frequency "baseband" analog and digital signal processing (Fig. 1.3). In other words, the baseband section is, in terms of the number of devices, yet several orders of magnitude more complex than the circuit of Fig. 1.2. The definition of RF and baseband will become clear later, but here we note that the RF section is still the design bottleneck of the entire system. This is so for three reasons.

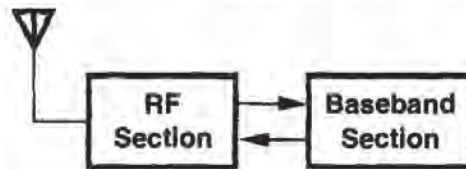


Figure 1.3 RF and baseband processing in a transceiver.

Multidisciplinary Field In contrast to other types of analog and mixed-signal circuits, RF systems demand a good understanding of many areas that are not directly related to integrated circuits (ICs). Shown in Fig. 1.4, most of these areas have been studied extensively for more than half a century, making it difficult for an IC designer to acquire the necessary knowledge in a reasonable amount of time. Even at present, the literature pertaining to RF design appears in more than 30 journals and conferences.

Owing to this issue, traditional wireless system design has been carried out at somewhat disjointed levels of abstraction: communication theorists create the modulation scheme and baseband signal processing; RF system experts plan the transceiver architecture; IC designers develop each of the building blocks; and manufacturers "glue" the ICs and other external components together. In fact, architectures are often planned according to the available off-the-shelf components, and ICs are designed to serve as many architectures as possible, leading to a great deal of redundancy at both system and circuit levels.

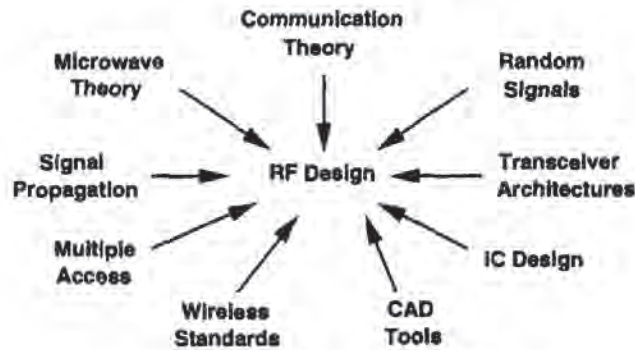


Figure 1.4 Disciplines required in RF design.

As the industry moves toward higher integration and lower cost, RF and wireless design demands increasingly more “concurrent engineering,” thereby requiring IC designers to have sufficient knowledge of all the disciplines shown in Fig. 1.4.

RF Design Hexagon RF circuits must process analog signals with a wide dynamic range at high frequencies. It is interesting to note that the signals must be treated as analog even if the modulation is digital or the amplitude carries no information. (This will be clarified in later chapters.) The trade-offs involved in the design of such circuits can be summarized in the “RF design hexagon” shown in Fig. 1.5, where almost any two of the six parameters trade with each other to some extent. We will see these trade-offs in Chapters 6 to 9, but our observation here is that, while digital circuits directly benefit from advances in IC technologies, RF circuits do not as much. This issue is exacerbated by the fact that RF circuits often require external components—for example, inductors—that are difficult to bring onto the chip even in modern IC processes.

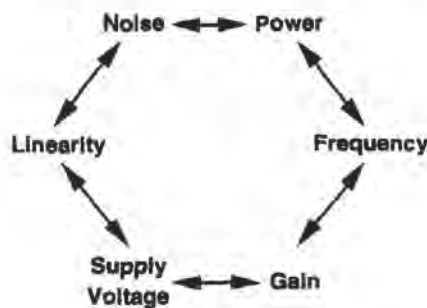


Figure 1.5 RF design hexagon.

Design Tools Computer-aided analysis and synthesis tools for RF ICs are still in their infancy, forcing the designer to rely on experience, intuition, or inefficient simulation techniques to predict the performance. For example, nonlinearity, time variance, and noise in RF circuits usually require studying the spectrum of signals, but the standard ac analysis available in SPICE uses

only linear, time-invariant models. Thus, circuits are simulated in the time domain so as to include nonlinear or time-variant effects, and the resulting waveforms are subsequently transformed to the frequency domain to obtain the spectrum. The difficulty is that the time-domain simulation must be run for a long period to resolve closely spaced frequency components. In addition, spectral averaging techniques may be necessary if random noise is used in the time-domain analysis.

Another issue in simulating RF circuits relates to external components that cannot be modeled by typical devices in SPICE. For example, surface acoustic wave (SAW) filters, used in both the receive and the transmit paths, exhibit input and output impedances that can be characterized only by scattering (S) parameters—essentially a table of numbers. Modeling such circuits with RLC networks provides a first-order approximation, but it may not predict effects such as instability and impedance mismatch.

1.3 APPLICATIONS

In addition to familiar wireless products such as pagers and cellular phones, RF technology has created many other markets that display a great potential for rapid growth, each presenting its own set of challenges to RF designers.

WLANs Communication among people or pieces of equipment in a crowded area can be realized through a wireless local area network (WLAN). Using frequency bands around 900 MHz and 2.4 GHz, WLAN transceivers can provide mobile connectivity in offices, hospitals, factories, etc., obviating the need for cumbersome wired networks. Portability and reconfigurability are prominent features of WLANs.

GPS The use of GPS to determine one's location as well as obtain directions becomes attractive to the consumer market as the cost and power dissipation of GPS receivers drop. Operating in the 1.5-GHz range, such systems are under consideration by automobile manufacturers, but they may be available as low-cost hand-held products sometime in the near future.

RF IDs RF identification systems, simply called "RF IDs," are small, low-cost tags that can be attached to objects or persons so as to track their position. Applications range from luggage in airports to troops in military operations. Low power consumption is especially critical here as the tag's lifetime may be determined by that of a single small battery. RF ID products in the 900-MHz and 2.4-GHz range have recently appeared in the market.

Home Satellite Network The programs and services available through satellite television have attracted many consumers to home satellite networks. Operating in the 10-GHz range, these networks require the addition of a dish antenna and a receiver to a television set and directly compete with cable TV.

1.4 ANALOG AND DIGITAL SYSTEMS

To gain a general, preliminary view of wireless transceivers, first consider the simple “analog” architecture shown in Fig. 1.6. In the transmit path, the signal generated by the microphone modulates a high-frequency carrier, and the result is amplified and “buffered” so as to drive the antenna. In the receive path, the signal is amplified by a low-noise amplifier (LNA), the spectrum is translated to a lower frequency by a “downconverter” (usually a mixer) to facilitate subsequent demodulation, and the demodulated output is amplified to drive the speaker.

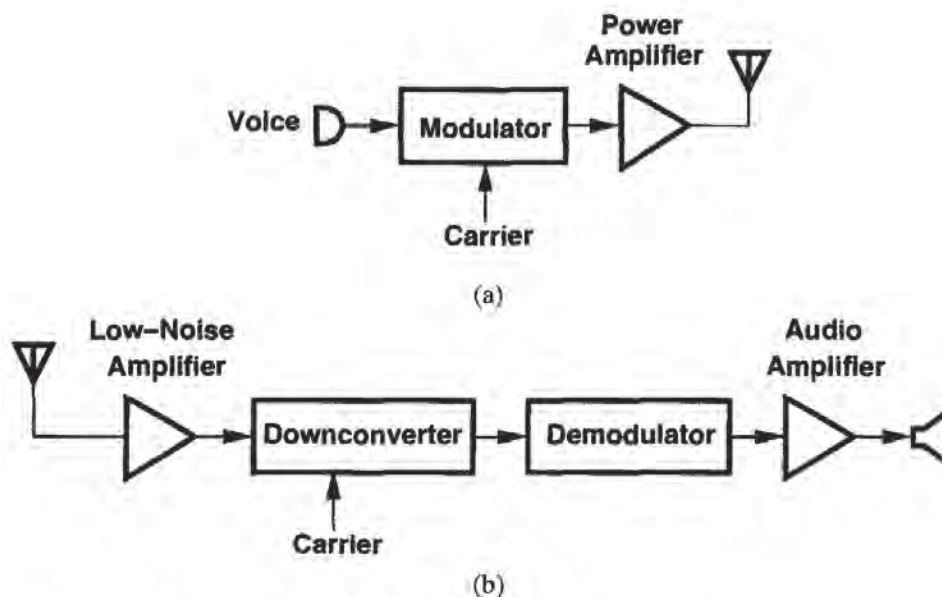


Figure 1.6 Block diagram of a generic analog RF system: (a) transmitter, (b) receiver.

Now, consider the “digital” transceiver shown in Fig. 1.7. Here, the voice signal is first digitized by an analog-to-digital converter (ADC) and compressed to reduce the bit rate and hence the required bandwidth [Fig. 1.7(a)]. Next, the data undergoes “coding” and “interleaving” [2]. These two functions format the data such that the receiver can detect and minimize errors by performing the reverse operations. Since rectangular pulses are usually not optimum for modulation, the data is then “shaped” before it is applied to the modulator and the power amplifier (PA). In the receive path [Fig. 1.7(b)], the signal is amplified, downconverted, and digitized. Subsequently, demodulation, equalization, decoding and deinterleaving, and decompression are performed in the digital domain. The resulting data is then converted to analog form by a digital-to-analog converter (DAC), amplified, and applied to the speaker.

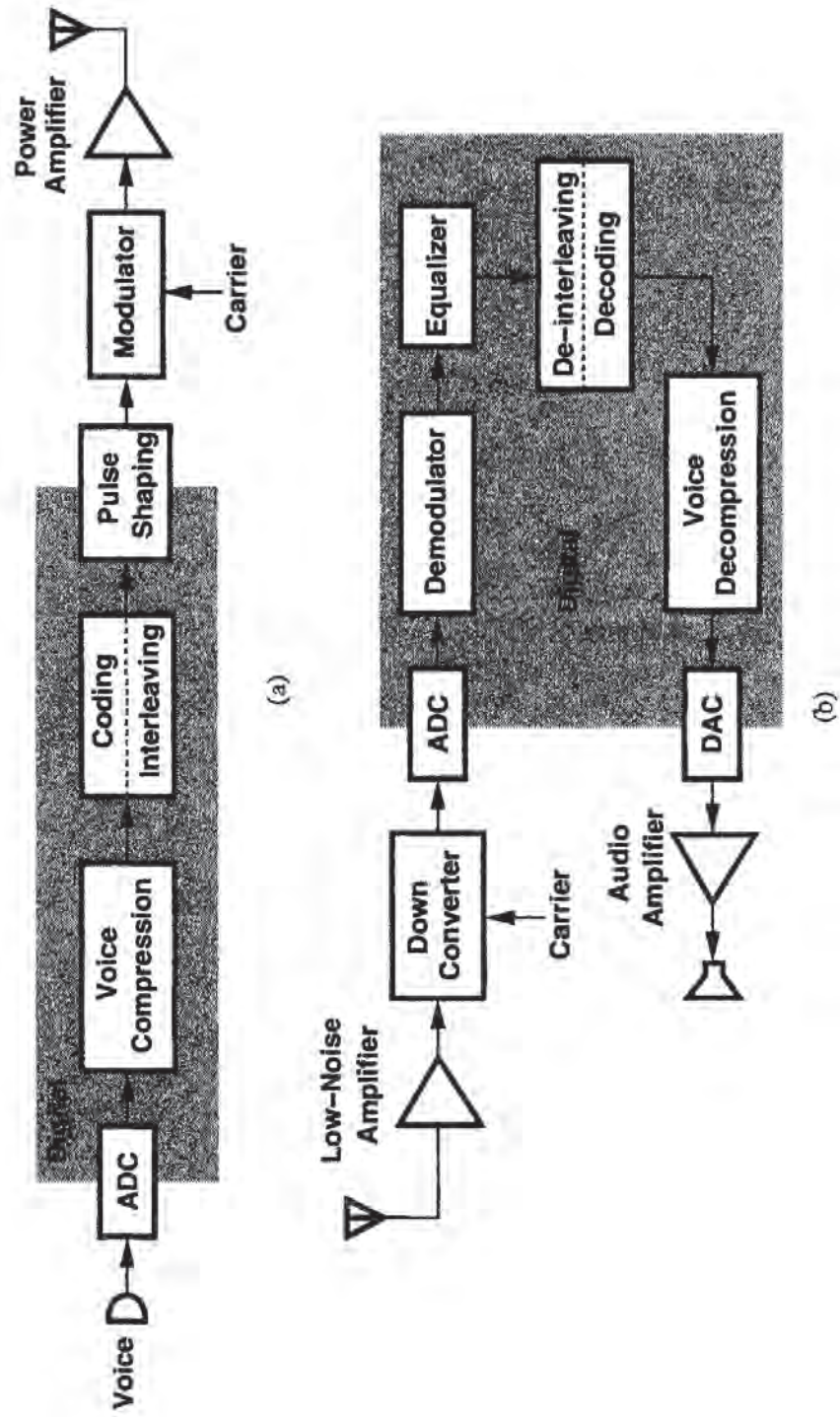


Figure 1.7 Block diagram of a generic digital RF system: (a) transmitter, (b) receiver.

In the simplest case, the performance of these transceivers can be quantified in terms of the maximum distance across which they operate while providing satisfactory reception. This is determined by both the power delivered to the antenna and the sensitivity of the receiver, particularly the noise of the LNA.¹ However, in a realistic environment, many different transceivers operate simultaneously, often in close proximity, thereby interfering with each other. Furthermore, the communication medium is quite complex: reflections from buildings and other obstacles can result in destructive interference at a given point, suppressing the received signal strength to undetectable levels. In such an environment, signal processing in the digital transceiver achieves a higher performance than that of the analog system.

Looking at the two transceivers of Figs. 1.6 and 1.7, the reader may wonder which parts are related to “RF electronics.” As we will see in other chapters, the exact definition of RF depends on the system, but we tentatively consider the signal to be in the RF domain if it is analog and its spectrum is not centered around zero frequency. This indicates that in Fig. 1.6 all the building blocks except for the audio amplifier are RF circuits. Similarly, in Fig. 1.7, the modulator, the PA, the LNA, and the downconverter operate in the RF range. Modern RF electronics, however, includes much more than the design of such building blocks. For example, the trade-offs at the system level must be carefully studied in each case [3, 4].

This book primarily deals with the design of RF integrated circuits, but, following the premise of Fig. 1.4, it provides the reader with background in other related areas as well. As will be seen throughout the book, RF ICs, RF architectures, and wireless systems are strongly coupled, and the reader will frequently need to utilize this background knowledge.

1.5 CHOICE OF TECHNOLOGY

The viable IC technology for RF circuits continues to change. Performance, cost, and time to market are three critical factors influencing the choice of technologies in the competitive RF industry. In addition, issues such as level of integration, form factor, and prior (successful) experience play an important role in the decisions made by the designers.

At present, GaAs and silicon bipolar and BiCMOS technologies constitute the major section of the RF market. Usually viewed as a low-yield, high-power, high-cost option, GaAs field-effect and heterojunction devices nonetheless have maintained a strong presence in RF products [5], especially in power amplifiers and front-end switches.

While GaAs processes offer useful features such as higher (breakdown voltage)(cutoff frequency) product, semi-insulating substrate, and high-quality inductors and capacitors, silicon devices in a VLSI technology can potentially

¹ The type of modulation is also an important factor, but we ignore it for the moment.

provide both higher levels of integration and lower overall cost, as demonstrated in complex circuits such as frequency synthesizers. In fact, all building blocks of typical transceivers are available in silicon bipolar technologies from many manufacturers.

The third contender is CMOS technology. Supported by the enormous momentum of the digital market, CMOS devices have achieved high transit frequencies, e.g., tens of gigahertz in the 0.35- μm generation. As we will see in this book, "RF CMOS" has suddenly become the topic of active research. CMOS technology must nevertheless resolve a number of practical issues: substrate coupling of signals that differ in amplitude by 100 dB, parameter variation with temperature and process, and device modeling for RF operation.

REFERENCES

- [1] C. Marshall et al., "2.7 V GSM Transceiver ICs With On-Chip Filtering," *ISSCC Dig. Tech. Papers*, pp. 148–149, February 1995.
- [2] B. Sklar, *Digital Communications*, Englewood Cliffs, NJ: Prentice Hall, 1988.
- [3] K. Feher, *Wireless Digital Communications*, Englewood Cliffs, NJ: Prentice Hall, 1995.
- [4] R. Steele, ed. *Mobile Radio Communications*, Piscataway, NJ: IEEE Press, 1992.
- [5] R. Schneiderman, "GaAs Continues to Gain in Wireless Applications," *Wireless Systems Design*, pp. 14–16, March 1997.

2

BASIC CONCEPTS IN RF DESIGN

RF designers draw upon many concepts that originate from the theory of signals and systems. In this chapter, we describe these concepts and define the terminology used in RF electronics so as to prepare the reader for the material in the following chapters.

Beginning with nonlinear systems, we describe effects such as harmonic distortion, gain compression, cross modulation, and intermodulation. We then briefly study intersymbol interference and Nyquist signaling, review random processes and noise, and introduce approaches to representing noise in circuits. Finally, we describe passive impedance transformation.

2.1 NONLINEARITY AND TIME VARIANCE

A system is linear if its output can be expressed as a linear combination (superposition) of responses to individual inputs. More accurately, if for inputs $x_1(t)$ and $x_2(t)$

$$x_1(t) \rightarrow y_1(t), \quad x_2(t) \rightarrow y_2(t), \quad (2.1)$$

where the arrow denotes the operation of the system, then

$$ax_1(t) + bx_2(t) \rightarrow ay_1(t) + by_2(t), \quad (2.2)$$

for all values of the constants a and b . Any system that does not satisfy this condition is nonlinear. Note that according to this definition, we consider a system nonlinear if it has nonzero initial conditions or finite “offsets.”

A system is time invariant if a time shift in its input results in the same time shift in its output. That is, if $x(t) \rightarrow y(t)$, then $x(t - \tau) \rightarrow y(t - \tau)$, for all values of τ . A system is time variant if it does not satisfy this condition.

While nonlinearity and time variance are intuitively obvious concepts, they may be confused with *each other* in some cases. For example, consider the switching circuit shown in Fig. 2.1(a). The control terminal of the switch is driven by $v_{in1}(t) = A_1 \cos \omega_1 t$ and the input terminal by $v_{in2}(t) = A_2 \cos \omega_2 t$. We assume the switch is on if $v_{in1} > 0$ and off otherwise. Is this system nonlinear or time variant? If, as shown in Fig. 2.1(b), the path of interest is from v_{in1} to v_{out} , (while v_{in2} is part of the system and still equal to $A_2 \cos \omega_2 t$), then the system is nonlinear because the control is only sensitive to the polarity of v_{in1} , and time variant because v_{out} also depends on v_{in2} . On the other hand, if, as shown in Fig. 2.1(c), the path of interest is from v_{in2} to v_{out} (while v_{in1} is part of the system and still equal to $A_1 \cos \omega_1 t$), then the system is *linear* [Eq. (2.2)] but time variant. Thus, general statements such as “switches are nonlinear” are ambiguous. As we will see in Chapter 6, these distinctions are critical in the design of mixers.

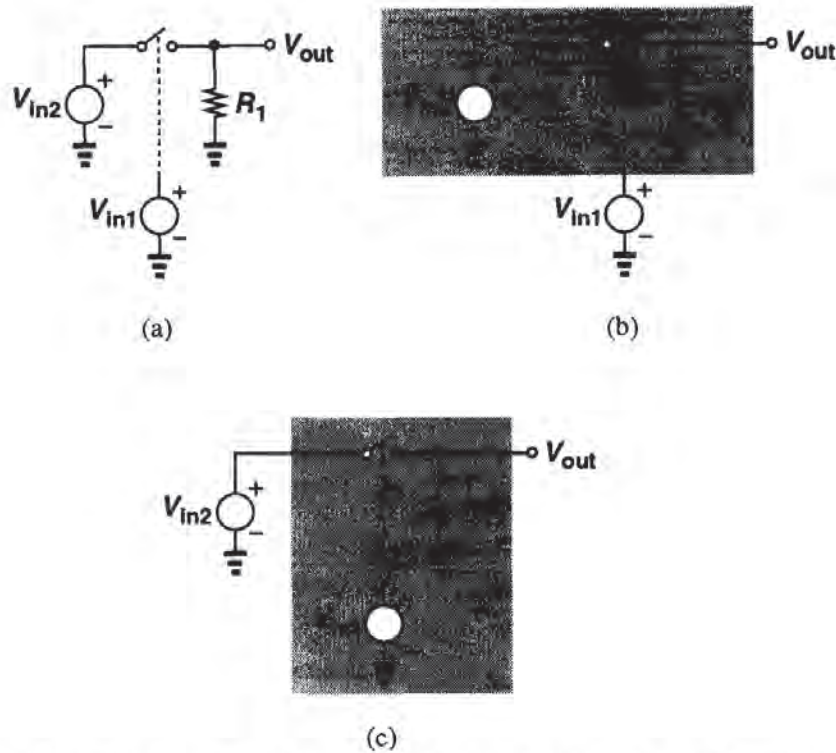


Figure 2.1 (a) Simple switching circuit, (b) nonlinear time-variant system, (c) linear time-variant system.

Another interesting result of the above observation is that a linear system *can* generate frequency components that do not exist in the input signal. This is possible if the system is time variant, for example, Fig. 2.1(c). Since in this

circuit, v_{out} can be considered as the product of v_{in2} and a square wave toggling between 0 and 1, the output spectrum is given by

$$V_{out}(f) = V_{in2}(f) * \sum_{n=-\infty}^{+\infty} \frac{\sin(n\pi/2)}{n\pi} \delta\left(f - \frac{n}{T_1}\right) \quad (2.3)$$

$$= \sum_{n=-\infty}^{+\infty} \frac{\sin(n\pi/2)}{n\pi} V_{in2}\left(f - \frac{n}{T_1}\right), \quad (2.4)$$

where $*$ denotes convolution, $\delta(\cdot)$ is the Dirac delta function, and $T_1 = 2\pi/\omega_1$. Thus, the output consists of vertically scaled replicas of $V_{in2}(f)$ shifted by n/T_1 .

A system is called “memoryless” if its output does not depend on the past values of its input. For a memoryless linear system,

$$y(t) = \alpha x(t), \quad (2.5)$$

where α is a function of time if the system is time variant [e.g., Fig. 2.1(c)]. For a memoryless nonlinear system, the input-output relationship can be approximated with a polynomial,

$$y(t) = \alpha_0 + \alpha_1 x(t) + \alpha_2 x^2(t) + \alpha_3 x^3(t) + \dots, \quad (2.6)$$

where α_j are in general functions of time if the system is time variant. Fig. 2.2(a) illustrates an example where the input signal is applied to the base of Q_1 while Q_2 and Q_3 are periodically switched by means of a square wave. For ideal bipolar transistors, the circuit can be viewed as in Fig. 2.2(b) and

$$v_{out}(t) = \left(I_{S1} \exp \frac{v_{in}}{V_T}\right) s(t) \cdot R, \quad (2.7)$$

where I_{S1} represents the saturation current of Q_1 , $V_T = kT/q$, and $s(t)$ is a square wave toggling between -1 and $+1$.

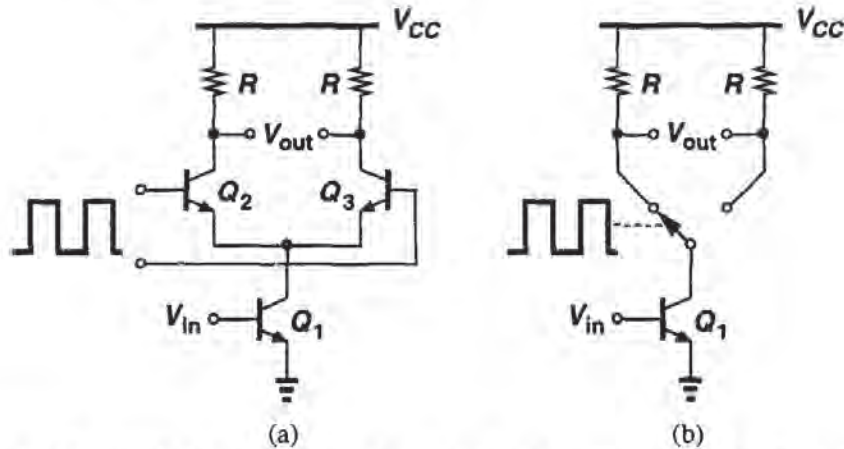


Figure 2.2 (a) Switching differential pair with tail current source driven by a signal, (b) equivalent circuit of (a).

The system described by (2.6) has “odd symmetry” if its response to $-x(t)$ is the negative of that to $x(t)$. This occurs if $\alpha_j = 0$ for even j . A circuit having odd symmetry is called differential or “balanced.” For example, the bipolar differential pair of Fig. 2.3 exhibits the following input-output characteristic:

$$v_{\text{out}} = RI_{EE} \tanh \frac{v_{\text{in}}}{2V_T}, \quad (2.8)$$

which is an odd function.

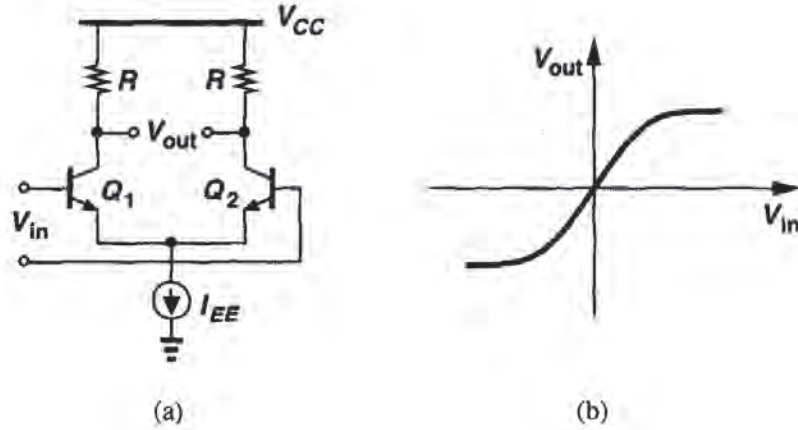


Figure 2.3 Bipolar differential pair along with its input-output characteristic.

A system is called “dynamic” if its output depends on the past values of its input(s) or output(s). For a linear, time-invariant, dynamic system,

$$y(t) = h(t) * x(t), \quad (2.9)$$

where $h(t)$ denotes the impulse response. If a dynamic system is linear but time variant, its impulse response depends on the time origin; if $\delta(t) \rightarrow h(t)$, then $\delta(t - \tau) \rightarrow h(t, \tau)$. Thus,

$$y(t) = h(t, \tau) * x(t). \quad (2.10)$$

Finally, if a system is both nonlinear and dynamic, then its impulse response can be approximated with a Volterra series [1, 2], a topic beyond the scope of this book.

2.1.1 Effects of Nonlinearity

While many analog and RF circuits can be approximated with a linear model to obtain their response to small signals, nonlinearities often lead to interesting and important phenomena. For simplicity, we limit our analysis to memoryless, time-variant systems and assume

$$y(t) \approx \alpha_1 x(t) + \alpha_2 x^2(t) + \alpha_3 x^3(t). \quad (2.11)$$

The reader is cautioned, however, that the effect of storage elements and higher-order nonlinear terms must be carefully examined to ensure (2.11) is a plausible representation.

Harmonics If a sinusoid is applied to a nonlinear system, the output generally exhibits frequency components that are integer multiples of the input frequency. In Eq. (2.11), if $x(t) = A \cos \omega t$, then

$$y(t) = \alpha_1 A \cos \omega t + \alpha_2 A^2 \cos^2 \omega t + \alpha_3 A^3 \cos^3 \omega t \quad (2.12)$$

$$= \alpha_1 A \cos \omega t + \frac{\alpha_2 A^2}{2} (1 + \cos 2\omega t) + \frac{\alpha_3 A^3}{4} (3 \cos \omega t + \cos 3\omega t) \quad (2.13)$$

$$= \frac{\alpha_2 A^2}{2} + \left(\alpha_1 A + \frac{3\alpha_3 A^3}{4} \right) \cos \omega t + \frac{\alpha_2 A^2}{2} \cos 2\omega t + \frac{\alpha_3 A^3}{4} \cos 3\omega t. \quad (2.14)$$

In Eq. (2.14), the term with the input frequency is called the “fundamental” and the higher-order terms the “harmonics.”

From the above expansion, we can make two observations. First, even-order harmonics result from α_j with even j and vanish if the system has odd symmetry, i.e., if it is fully differential. In reality, however, mismatches corrupt the symmetry, yielding finite even-order harmonics. Second, in (2.14) the amplitude of the n th harmonic consists of a term proportional to A^n and other terms proportional to higher powers of A . Neglecting the latter for small A , we can assume the n th harmonic grows approximately in proportion to A^n .

Gain Compression The small-signal gain of a circuit is usually obtained with the assumption that harmonics are negligible. For example, if in (2.14), $\alpha_1 A$ is much greater than all the other factors that contain A , then the small-signal gain is equal to α_1 . This quantity can be seen in the familiar differential pair of Fig. 2.3 to be equal to

$$\frac{v_{\text{out}}}{v_{\text{in}}} = \frac{I_{EE} R}{2V_T}. \quad (2.15)$$

However, as the signal amplitude increases, the gain begins to vary. In fact, nonlinearity can be viewed as variation of the small-signal gain with the input level. This is evident from the term $3\alpha_3 A^3/4$ added to $\alpha_1 A$ in (2.14), as well as the input-output characteristic shown in Fig. 2.3.

In most circuits of interest, the output is a “compressive” or “saturating” function of the input; that is, the gain approaches zero for sufficiently high input levels. In (2.14) this occurs if $\alpha_3 < 0$. Written as $\alpha_1 + 3\alpha_3 A^2/4$, the gain is therefore a decreasing function of A . In RF circuits, this effect is quantified by the “1-dB compression point,” defined as the input signal level that causes the

small-signal gain to drop by 1 dB. If plotted on a log-log scale as a function of the input level, the output level falls below its ideal value by 1 dB at the 1-dB compression point (Fig. 2.4).

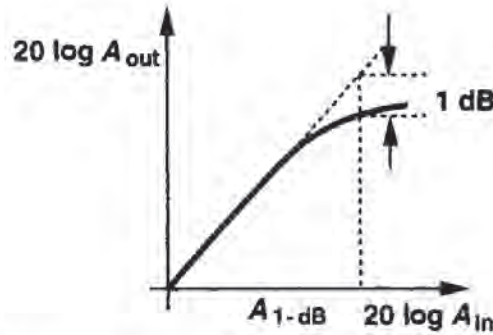


Figure 2.4 Definition of the 1-dB compression point.

To calculate the 1-dB compression point, we can write from (2.14)

$$20 \log |\alpha_1 + \frac{3}{4} \alpha_3 A_{1-dB}^2| = 20 \log |\alpha_1| - 1 \text{ dB}. \quad (2.16)$$

That is,

$$A_{1-dB} = \sqrt{0.145 \left| \frac{\alpha_1}{\alpha_3} \right|}. \quad (2.17)$$

A measure of the maximum input range of the circuit, the 1-dB compression point occurs around -20 to -25 dBm (63.2 to 35.6 mV_{pp} in a $50\text{-}\Omega$ system) in typical front-end RF amplifiers.

Desensitization and Blocking Circuits with compressive characteristics exhibit an interesting effect when they process a weak, desired signal along with a strong interferer. Since a large signal tends to reduce the “average” gain of the circuit, the weak signal may experience a vanishingly small gain. Called “desensitization,” this effect can be analyzed for the characteristics of (2.11) by assuming $x(t) = A_1 \cos \omega_1 t + A_2 \cos \omega_2 t$. The output is

$$y(t) = \left(\alpha_1 A_1 + \frac{3}{4} \alpha_3 A_1^3 + \frac{3}{2} \alpha_3 A_1 A_2^2 \right) \cos \omega_1 t + \dots, \quad (2.18)$$

which, for $A_1 \ll A_2$, reduces to

$$y(t) = \left(\alpha_1 + \frac{3}{2} \alpha_3 A_2^2 \right) A_1 \cos \omega_1 t + \dots. \quad (2.19)$$

Thus, the gain for the desired signal is equal to $(\alpha_1 + 3\alpha_3 A_2^2/2)$, a decreasing function of A_2 if $\alpha_3 < 0$. For sufficiently large A_2 , the gain drops to zero, and we say the signal is “blocked.” In RF design, the term “blocking signal” usually

refers to interferers that desensitize a circuit even if the gain does not fall to zero. Many RF receivers must be able to withstand blocking signals 60 to 70 dB greater than the wanted signal.

Cross Modulation Another phenomenon that occurs when a weak signal and a strong interferer pass through a nonlinear system is the transfer of modulation (or noise) on the amplitude of the interferer to the amplitude of the weak signal. Called “cross modulation,” this effect is evident from Eq. (2.19), where variations in A_2 affect the amplitude of the output component at ω_1 . For example, if the amplitude of the interferer is modulated by a sinusoid $A_2(1 + m \cos \omega_m t) \cos \omega_2 t$, where m is the modulation index and less than unity, then (2.19) assumes the following form:

$$y(t) = \left[\alpha_1 A_1 + \frac{3}{2} \alpha_3 A_1 A_2^2 \left(1 + \frac{m^2}{2} + \frac{m^2}{2} \cos 2\omega_m t + 2m \cos \omega_m t \right) \right] \cos \omega_1 t + \dots \quad (2.20)$$

Thus, the desired signal at the output contains amplitude modulation at ω_m and $2\omega_m$.

A common case of cross modulation arises in amplifiers that must *simultaneously* process many independent signal channels, e.g., in cable television transmitters. Nonlinearities in the amplifier corrupt each signal with the amplitude variations in other channels.

Intermodulation While harmonic distortion is often used to describe nonlinearities of analog circuits, certain cases require other measures of nonlinear behavior. For example, suppose the nonlinearity of an active low-pass filter is to be evaluated. If, as depicted in Fig. 2.5, the input sinusoid frequency is chosen such that its harmonics fall out of the passband, then the output distortion appears quite small even if the input stage of the filter introduces substantial nonlinearity. Thus, another type of test is required here. Commonly used is the “intermodulation distortion” in a “two-tone” test.

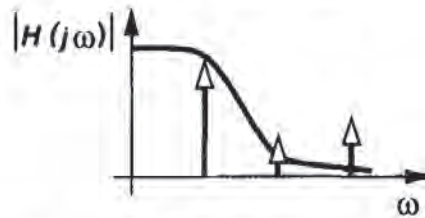


Figure 2.5 Harmonic distortion in a low-pass filter.

When two signals with different frequencies are applied to a nonlinear system, the output in general exhibits some components that are not harmonics of the input frequencies. Called intermodulation (IM), this phenomenon arises

from “mixing” (multiplication) of the two signals when their sum is raised to a power greater than unity. To understand how Eq. (2.11) leads to intermodulation, assume $x(t) = A_1 \cos \omega_1 t + A_2 \cos \omega_2 t$. Thus,

$$y(t) = \alpha_1(A_1 \cos \omega_1 t + A_2 \cos \omega_2 t) + \alpha_2(A_1 \cos \omega_1 t + A_2 \cos \omega_2 t)^2 + \alpha_3(A_1 \cos \omega_1 t + A_2 \cos \omega_2 t)^3. \quad (2.21)$$

Expanding the left side and discarding dc terms and harmonics, we obtain the following intermodulation products:

$$\omega = \omega_1 \pm \omega_2 : \alpha_2 A_1 A_2 \cos(\omega_1 + \omega_2)t + \alpha_2 A_1 A_2 \cos(\omega_1 - \omega_2)t \quad (2.22)$$

$$= 2\omega_1 \pm \omega_2 : \frac{3\alpha_3 A_1^2 A_2}{4} \cos(2\omega_1 + \omega_2)t + \frac{3\alpha_3 A_1^2 A_2}{4} \cos(2\omega_1 - \omega_2)t \quad (2.23)$$

$$= 2\omega_2 \pm \omega_1 : \frac{3\alpha_3 A_2^2 A_1}{4} \cos(2\omega_2 + \omega_1)t + \frac{3\alpha_3 A_2^2 A_1}{4} \cos(2\omega_2 - \omega_1)t \quad (2.24)$$

and these fundamental components

$$\begin{aligned} \omega = \omega_1, \omega_2 : & \left(\alpha_1 A_1 + \frac{3}{4}\alpha_3 A_1^3 + \frac{3}{2}\alpha_3 A_1 A_2^2 \right) \cos \omega_1 t \\ & + \left(\alpha_1 A_2 + \frac{3}{4}\alpha_3 A_2^3 + \frac{3}{2}\alpha_3 A_2 A_1^2 \right) \cos \omega_2 t. \end{aligned} \quad (2.25)$$

Of particular interest are the third-order IM products at $2\omega_1 - \omega_2$ and $2\omega_2 - \omega_1$, illustrated in Fig. 2.6. The key point here is that if the difference between ω_1 and ω_2 is small, the components at $2\omega_1 - \omega_2$ and $2\omega_2 - \omega_1$ appear in the vicinity of ω_1 and ω_2 , thus revealing nonlinearities even in cases such as the LPF of Fig. 2.5. In a typical two-tone test, $A_1 = A_2 = A$, and the ratio of the amplitude of the output third-order products to $\alpha_1 A$ defines the IM distortion. For example, if $\alpha_1 A = 1 \text{ V}_{pp}$, and $3\alpha_3 A^3/4 = 10 \text{ mV}_{pp}$, then we say the IM components are at -40 dBc , where the letter “c” means “with respect to the carrier.”

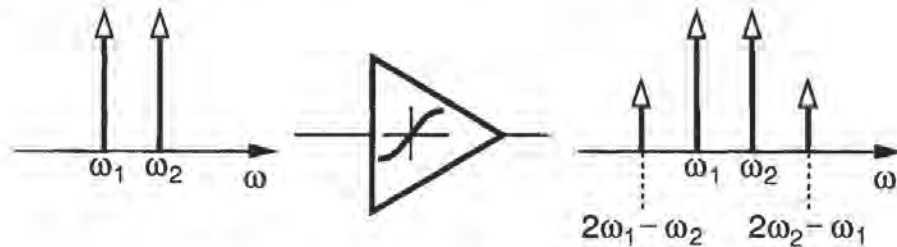


Figure 2.6 Intermodulation in a nonlinear system.

Intermodulation is a troublesome effect in RF systems. As shown in Fig. 2.7, if a weak signal accompanied by two strong interferers experiences third-order nonlinearity, then one of the IM products falls in the band of interest, corrupting the desired component. While operating on the *amplitude* of the signals, this effect degrades the performance even if the modulation is on the phase (because zero-crossing points are still affected.) Note that this phenomenon cannot be directly quantified by *harmonic* distortion.

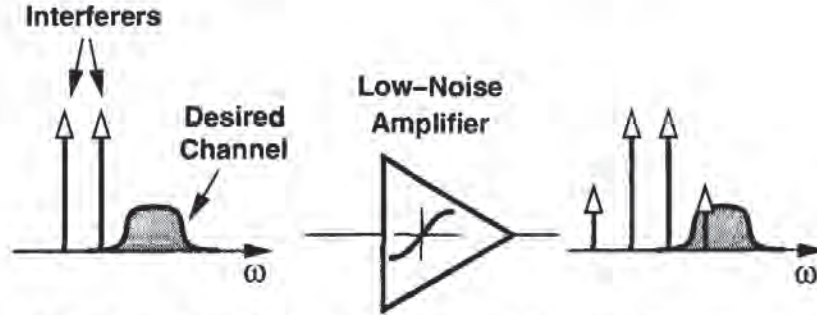


Figure 2.7 Corruption of a signal due to intermodulation between two interferers.

The corruption of signals due to third-order intermodulation of two nearby interferers is so common and so critical that a performance metric has been defined to characterize this behavior. Called the “third intercept point” (IP_3), this parameter is measured by a two-tone test in which A is chosen to be sufficiently small so that higher-order nonlinear terms are negligible and the gain is relatively constant and equal to α_1 . From (2.23), (2.24), and (2.25), we note that as A increases, the fundamentals increase in proportion to A , whereas the third-order IM products increase in proportion to A^3 [Fig. 2.8(a)]. Plotted on a logarithmic scale [Fig. 2.8(b)], the magnitude of the IM products grows at three times the rate at which the main components increase. The third-order intercept point is defined to be at the intersection of the two lines. The horizontal coordinate of this point is called the input IP_3 (IIP_3), and the vertical coordinate is called the output IP_3 (OIP_3).

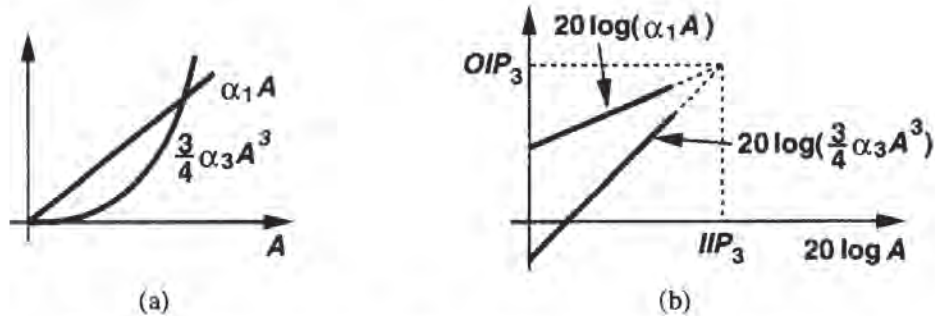


Figure 2.8 Growth of output components in an intermodulation test.

It is important to appreciate the advantage of IP_3 over a simple IM measurement. If the magnitude of IM products (normalized to that of the carrier) is used as a measure of linearity, then the input amplitude with which the test is performed must be specified. The third intercept point, on the other hand, is a unique quantity that by itself can serve as a means of comparing the linearity of different circuits.

From the input-output characteristic of Eq. (2.11), we can derive a simple expression for IP_3 . Let $x(t) = A \cos \omega_1 t + A \cos \omega_2 t$. Then,

$$\begin{aligned} y(t) = & \left(\alpha_1 + \frac{9}{4} \alpha_3 A^2 \right) A \cos \omega_1 t + \left(\alpha_1 + \frac{9}{4} \alpha_3 A^2 \right) A \cos \omega_2 t \\ & + \frac{3}{4} \alpha_3 A^3 \cos(2\omega_1 - \omega_2)t + \frac{3}{4} \alpha_3 A^3 \cos(2\omega_2 - \omega_1)t + \dots \end{aligned} \quad (2.26)$$

If $\alpha_1 \gg 9\alpha_3 A^2/4$, the input level for which the output components at ω_1 and ω_2 have the same amplitude as those at $2\omega_1 - \omega_2$ and $2\omega_2 - \omega_1$ is given by

$$|\alpha_1| A_{IP3} = \frac{3}{4} |\alpha_3| A_{IP3}^3. \quad (2.27)$$

Thus, the input IP_3 is

$$A_{IP3} = \sqrt{\frac{4}{3} \frac{|\alpha_1|}{|\alpha_3|}}, \quad (2.28)$$

and the output IP_3 is equal to $\alpha_1 A_{IP3}$.

The parameter IP_3 characterizes only third-order nonlinearities. In practice, if the input level is increased to reach the intercept point, the assumption $\alpha_1 \gg 9\alpha_3 A^2/4$ no longer holds, the gain drops, and higher-order IM products become significant. In fact, in many circuits the IP_3 is beyond the allowable input range, sometimes even higher than the supply voltage. Thus, the practical method of obtaining the IP_3 is to measure the characteristic of Fig. 2.8(b) for small input amplitudes and use linear extrapolation on a logarithmic scale to find the intercept point.

A quick method of measuring the IP_3 is as follows. Let us denote the input level at each frequency by A_{in} , the amplitude of the output components at ω_1 and ω_2 by A_{ω_1, ω_2} , and the amplitude of the IM_3 products by A_{IM3} . Then from (2.26), we have

$$\frac{A_{\omega_1, \omega_2}}{A_{IM3}} \approx \frac{|\alpha_1| A_{in}}{3 |\alpha_3| A_{in}^3 / 4} \quad (2.29)$$

$$= \frac{4 |\alpha_1|}{3 |\alpha_3|} \frac{1}{A_{in}^2}, \quad (2.30)$$

which, in conjunction with (2.28), reduces to

$$\frac{A_{\omega_1, \omega_2}}{A_{IM3}} = \frac{A_{IP3}^2}{A_{in}^2}. \quad (2.31)$$

Consequently,

$$20 \log A_{\omega_1, \omega_2} - 20 \log A_{IM3} = 20 \log A_{IP3}^2 - 20 \log A_{in}^2, \quad (2.32)$$

and

$$20 \log A_{IP3} = \frac{1}{2}(20 \log A_{\omega_1, \omega_2} - 20 \log A_{IM3}) + 20 \log A_{in}. \quad (2.33)$$

Thus, if all the signal levels are expressed in dBm, the *input* third intercept point is equal to half the difference between the magnitudes of the fundamentals and the IM_3 products at the *output* plus the corresponding input level [Fig. 2.9(a)]. The key point here is that IP_3 can be measured with only one input level, obviating the need for extrapolation.

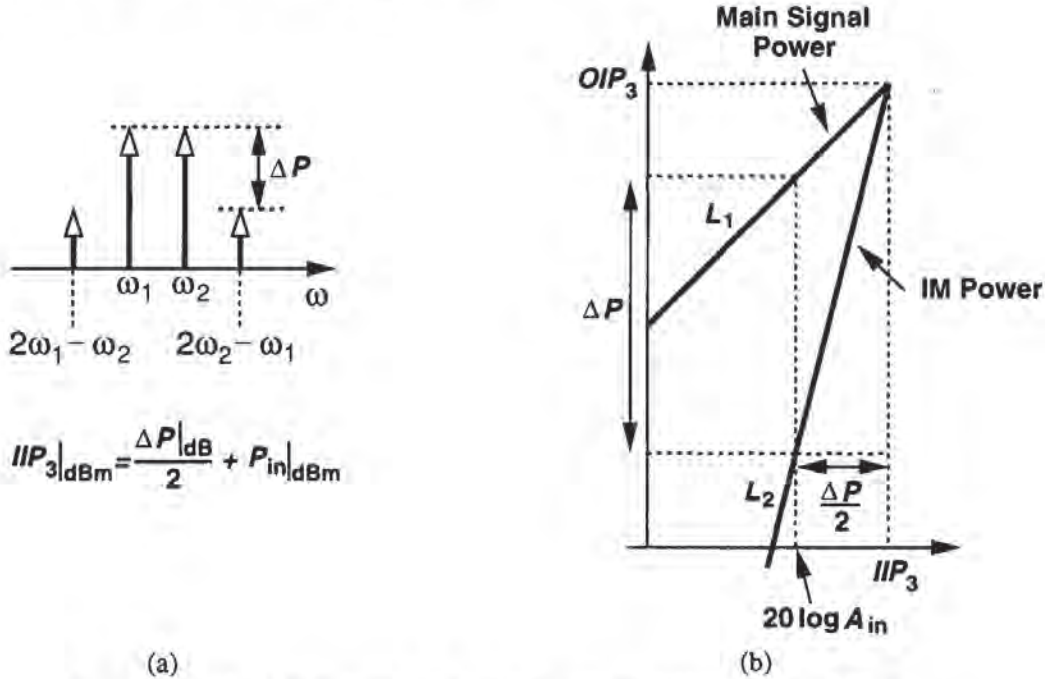


Figure 2.9 (a) Calculation of IP_3 without extrapolation, (b) graphical interpretation of (a).

Shown in Fig. 2.9(b) is a geometric interpretation of the above relationship. Since line L_1 has a slope equal to unity and line L_2 a slope equal to 3, an input increment $\Delta P/2$ yields an equal increment in L_1 and an increment equal to $3\Delta P/2$ in L_2 , reducing the difference between the two lines to zero.

The above approach provides an estimate of IP_3 in initial phases of the design or characterization. The actual value of IP_3 , however, must still be obtained through accurate extrapolation to ensure that all nonlinear and frequency-dependent effects are taken into account.

Another measurement method encountered in the literature is to apply a *single* tone, plot the third harmonic magnitude versus the input level, and obtain the intercept point by extrapolation. From the example of Fig. 2.5, we note that this technique does not yield a correct value for IP_3 .

To gain a better feeling about the required linearity in typical RF systems, let us calculate the amount of corruption that a $1\text{-}\mu\text{V}_{rms}$ signal experiences by two 1-mV_{rms} interferers in an amplifier having an IP_3 of 70 mV_{rms} ($\approx -10\text{ dBm}$) (Fig. 2.10). Neglecting desensitization and cross modulation, we can write

$$\frac{A_{sig,out}}{A_{sig,in}} \approx \frac{A_{int,out}}{A_{int,in}}, \quad (2.34)$$

where A_{sig} denotes the signal amplitude and A_{int} the interferer amplitude. It follows from (2.31) that

$$\frac{A_{sig,out}}{A_{IM3,out}} = \frac{A_{sig,in} \cdot A_{IP3}^2}{A_{int,in}^3} \quad (2.35)$$

where $A_{sig,in} = 1\text{ }\mu\text{V}_{rms}$, $A_{IP3} = 70\text{ mV}_{rms}$, and $A_{int,in} = 1\text{ mV}_{rms}$. Thus, the ratio is equal to $4.9 \approx 13.8\text{ dB}$.

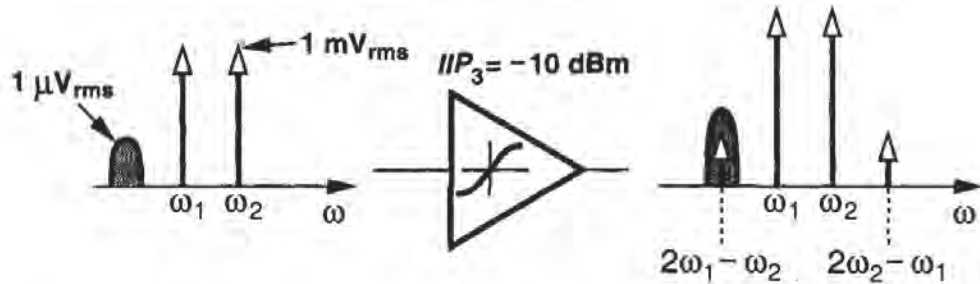


Figure 2.10 Example of achievable SNR in the presence of large interferers.

It is also instructive to find the relationship between the 1-dB compression point and the input IP_3 for a third-order nonlinearity. From (2.17) and (2.28), we conclude that the two are related by

$$\frac{A_{1\text{-dB}}}{A_{IP3}} = \frac{\sqrt{0.145}}{\sqrt{4/3}} \quad (2.36)$$

$$\approx -9.6\text{ dB}. \quad (2.37)$$

2.1.2 Cascaded Nonlinear Stages

Since in RF systems, signals are processed by cascaded stages, it is important to know how the nonlinearity of each stage is referred to the input of the cascade. In particular, it is desirable to calculate an overall input third intercept point in terms of the IP_3 and gain of the individual stages.

Consider two nonlinear stages in cascade (Fig. 2.11). If the input-output characteristics of the two stages are expressed, respectively, as

$$y_1(t) = \alpha_1 x(t) + \alpha_2 x^2(t) + \alpha_3 x^3(t) \quad (2.38)$$

$$y_2(t) = \beta_1 y_1(t) + \beta_2 y_1^2(t) + \beta_3 y_1^3(t), \quad (2.39)$$

then

$$\begin{aligned} y_2(t) = & \beta_1 [\alpha_1 x(t) + \alpha_2 x^2(t) + \alpha_3 x^3(t)] \\ & + \beta_2 [\alpha_1 x(t) + \alpha_2 x^2(t) + \alpha_3 x^3(t)]^2 \\ & + \beta_3 [\alpha_1 x(t) + \alpha_2 x^2(t) + \alpha_3 x^3(t)]^3. \end{aligned} \quad (2.40)$$

Considering only the first- and third-order terms, we have

$$y_2(t) = \alpha_1 \beta_1 x(t) + (\alpha_3 \beta_1 + 2\alpha_1 \alpha_2 \beta_2 + \alpha_1^3 \beta_3) x^3(t) + \dots \quad (2.41)$$

Thus, from (2.28)

$$A_{IP3} = \sqrt{\frac{4}{3} \left| \frac{\alpha_1 \beta_1}{\alpha_3 \beta_1 + 2\alpha_1 \alpha_2 \beta_2 + \alpha_1^3 \beta_3} \right|}. \quad (2.42)$$

Interestingly, proper choice of the values and signs of the terms in the denominator can yield an arbitrarily high I_{P3} . In practice, however, other considerations such as noise, gain, and active device characteristics may not permit this choice. As a worst-case estimate, we add the absolute values of the three terms in the denominator.

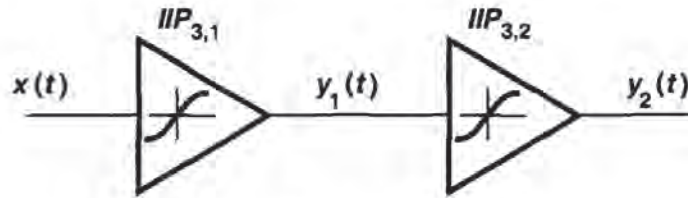


Figure 2.11 Cascaded nonlinear stages.

Equation (2.42) can be simplified if the two sides are inverted and squared:

$$\frac{1}{A_{IP3}^2} = \frac{3}{4} \frac{|\alpha_3 \beta_1| + |2\alpha_1 \alpha_2 \beta_2| + |\alpha_1^3 \beta_3|}{|\alpha_1 \beta_1|} \quad (2.43)$$

$$= \frac{1}{A_{IP3,1}^2} + \frac{3\alpha_2 \beta_2}{2\beta_1} + \frac{\alpha_1^2}{A_{IP3,2}^2}, \quad (2.44)$$

where $A_{IP3,1}$ and $A_{IP3,2}$ represent the input I_{P3} points of the first and second stages, respectively. Note that A_{IP3} , $A_{IP3,1}$, and $A_{IP3,2}$ are voltage quantities rather than power quantities.

From the above result, we note that as α_1 increases, the overall IP_3 decreases. This is because with higher gain in the first stage, the second stage senses larger input levels, thereby producing much greater IM_3 products. (Recall that IM_3 products grow with the third power of the input amplitude.)

To gain more insight, we assume $x(t) = A \cos \omega_1 t + A \cos \omega_2 t$ and identify the various IM products. Referring to Fig. 2.12, we make the following observations.¹ (1) The fundamental input components are amplified by approximately α_1 in the first stage and β_1 in the second. Thus the output fundamentals are $\alpha_1 \beta_1 A (\cos \omega_1 t + \cos \omega_2 t)$. (2) The IM_3 products generated by the first stage, namely, $(3\alpha_3/4)A^3 [\cos(2\omega_1 - \omega_2)t + \cos(2\omega_2 - \omega_1)t]$, are also amplified by β_1 when they appear at the output of the second stage. (3) The second stage senses $\alpha_1 A (\cos \omega_1 t + \cos \omega_2 t)$ at its input and hence generates these IM_3 products: $(3\beta_3/4)(\alpha_1 A)^3 \cos(2\omega_1 - \omega_2)t + (3\beta_3/4)(\alpha_1 A)^3 \cos(2\omega_2 - \omega_1)t$. (4) The second-order nonlinearity in $y_1(t)$ generates components at $\omega_1 - \omega_2$, $2\omega_1$, and $2\omega_2$. Upon experiencing a similar nonlinearity in the second stage, such components are translated to $2\omega_1 - \omega_2$ and $2\omega_2 - \omega_1$. More specifically, as shown in Fig. 2.12, $y_2(t)$ contains terms such as $2\beta_2[\alpha_1 A \cos \omega_1 t \cdot \alpha_2 A^2 \cos(\omega_1 - \omega_2)t]$ and $2\beta_2(\alpha_1 A \cos \omega_2 t \cdot 0.5\alpha_2 A^2 \cos 2\omega_1 t)$. The resulting third-order IM products can be expressed as $(3\alpha_1\alpha_2\beta_2 A^3/2)[\cos(2\omega_1 - \omega_2)t + \cos(2\omega_2 - \omega_1)t]$.

From these observations, we can write

$$\begin{aligned} y_2(t) = & \alpha_1 \beta_1 A (\cos \omega_1 t + \cos \omega_2 t) \\ & + \left(\frac{3\alpha_3\beta_1}{4} + \frac{3\alpha_1^3\beta_3}{4} + \frac{3\alpha_1\alpha_2\beta_2}{2} \right) A^3 [\cos(2\omega_1 - \omega_2)t \\ & + \cos(2\omega_2 - \omega_1)t] + \dots, \end{aligned} \quad (2.45)$$

obtaining the same IP_3 as above.

In many RF systems, each stage in a cascade has a narrow frequency band. Thus, the components described in the fourth observation above fall out of the band and are heavily attenuated. Consequently, the second term on the right-hand side of (2.44) becomes negligible, giving

$$\frac{1}{A_{IP3}^2} \approx \frac{1}{A_{IP3,1}^2} + \frac{\alpha_1^2}{A_{IP3,2}^2}. \quad (2.46)$$

This equation readily gives a general expression for three or more stages:

$$\frac{1}{A_{IP3}^2} \approx \frac{1}{A_{IP3,1}^2} + \frac{\alpha_1^2}{A_{IP3,2}^2} + \frac{\alpha_1^2\beta_1^2}{A_{IP3,3}^2} + \dots, \quad (2.47)$$

where $A_{IP3,3}$ denotes the input IP_3 of the third stage. Thus, if each stage in a cascade has a gain greater than unity, the nonlinearity of the latter stages becomes increasingly more critical because the IP_3 of each stage is effectively

¹ The spectrum of $A \cos \omega t$ consists of two impulses, each with a weight $A/2$. We drop the factor $1/2$ in the figures for simplicity. The end result is still correct.

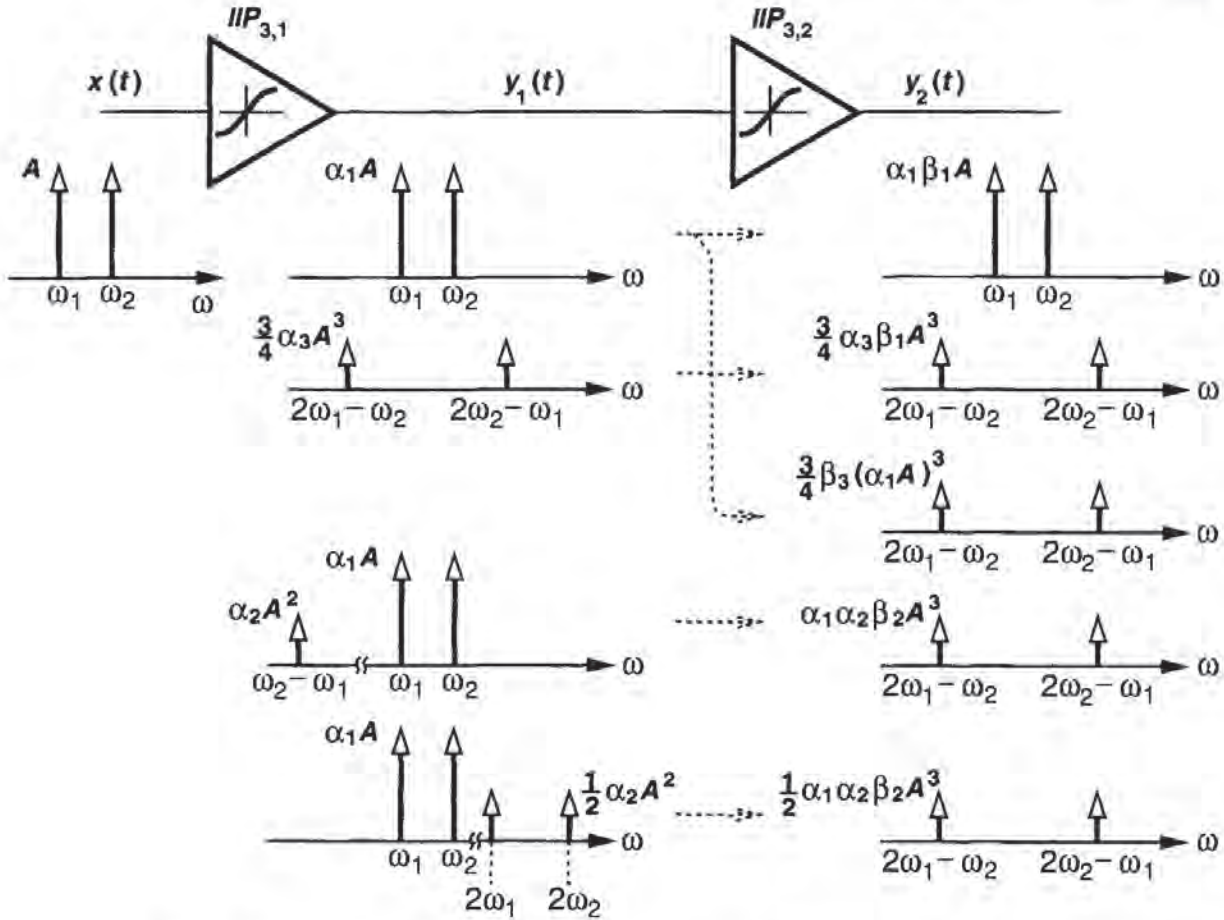


Figure 2.12 Intermodulation mechanisms in cascade of two nonlinear stages.

scaled *down* by the total gain preceding that stage. We should emphasize that (2.47) is merely an approximation. In practice, more precise calculations or simulations must be performed to predict the overall IP_3 .

2.2 INTERSYMBOL INTERFERENCE

Linear time-invariant systems can also “distort” a signal if they do not have sufficient bandwidth. Attenuation of high-frequency components of a periodic square wave in a low-pass filter is a familiar example of such behavior [Fig. 2.13(a)]. However, limited bandwidth has a more detrimental effect on *random* bit streams. To understand the issue, first recall that if a single ideal rectangular pulse is applied to a low-pass filter, the output exhibits an exponential tail that becomes more significant as the filter bandwidth decreases. This occurs fundamentally because a signal cannot be both time limited and bandwidth limited: when the time-limited pulse passes through the band-limited system, the output must extend to infinity in the time domain.

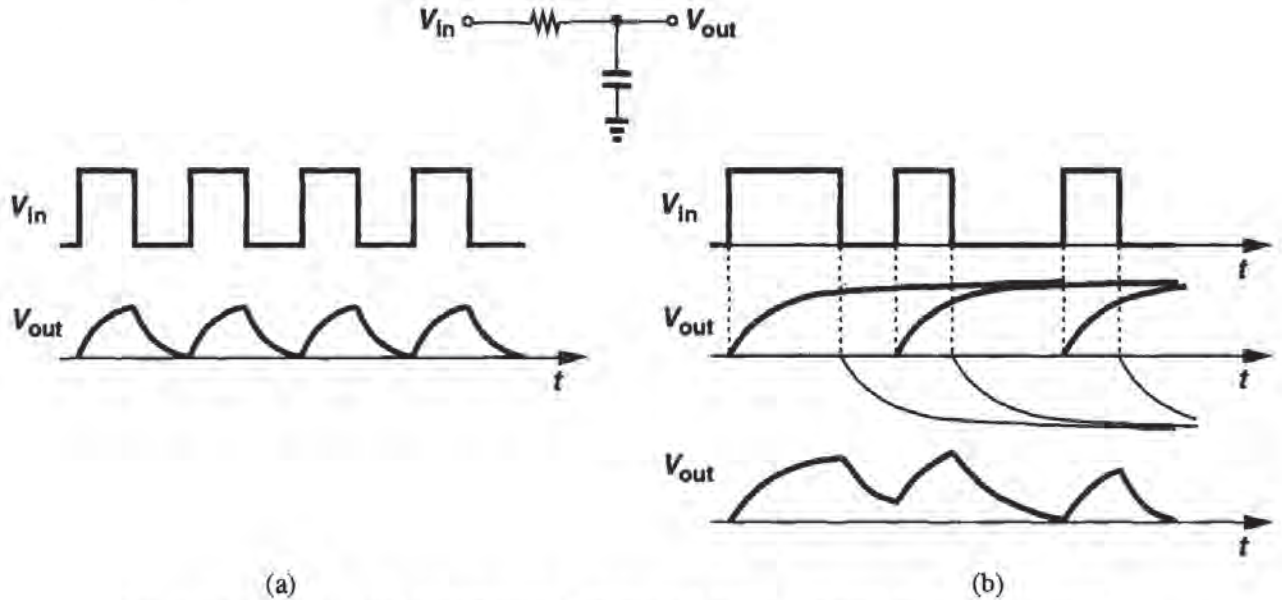


Figure 2.13 Response of a low-pass filter to (a) a periodic square wave, (b) a random sequence of ONES and ZEROS.

Now suppose the output of a digital system consists of a random sequence of ONES and ZEROS, each ONE represented by an ideal rectangular pulse and each ZERO by the absence of such a pulse. If this sequence is applied to a low-pass filter, the output can be obtained as the superposition of the responses to each input bit [Fig. 2.13(b)]. We note that each bit level is corrupted by decaying tails created by previous bits. Called “intersymbol interference” (ISI), this phenomenon leads to higher error rate in the detection of random waveforms that are transmitted through band-limited channels.

The problem of ISI is particularly troublesome in wireless communications because the bandwidth allocated to each channel is fairly narrow. Methods of reducing ISI include pulse shaping (“Nyquist signaling”) in the transmitter and “equalization” in the receiver. We briefly describe Nyquist signaling here and refer the reader to the extensive literature on equalization for ISI mitigation [3, 4].

In order to reduce ISI, the pulse shape can be chosen such that it is less susceptible to interference with its shifted replicas. In Nyquist signaling, each pulse is allowed to overlap with past and future pulses, but the shape is selected such that ISI is zero at certain points in time. Illustrated in Fig. 2.14, the idea is that all other pulses go through zero at the point when the present pulse reaches its peak. Thus, if the bit stream is sampled at $t = kT_S$, no ISI exists.

A simple calculation leads to a basic condition for Nyquist signals. For a pulse shape, $p(t)$, to introduce zero ISI, we have

$$p(kT_S) = 1 \quad \text{if } k = 0 \quad (2.48)$$

$$= 0 \quad \text{if } k \neq 0 \quad (2.49)$$

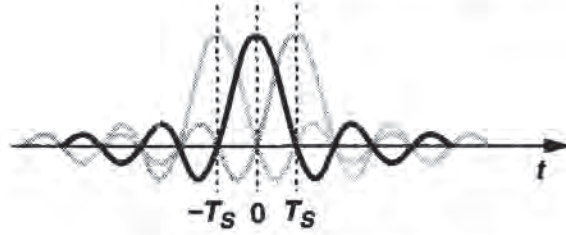


Figure 2.14 Pulse shape with no ISI.

Using a train of impulses to sample this pulse, we obtain

$$p(t) \cdot \sum \delta(t - kT_s) = \delta(t). \quad (2.50)$$

Taking the Fourier transform of both sides, we have

$$P(f) * \frac{1}{T_s} \sum \delta\left(f - \frac{k}{T_s}\right) = 1. \quad (2.51)$$

That is,

$$\frac{1}{T_s} \sum P\left(f - \frac{k}{T_s}\right) = 1. \quad (2.52)$$

Originally proposed by Nyquist and shown in Fig. 2.15, this result indicates that the shifted replicas of $P(f)$ must add up to a flat spectrum. For example, a sinc waveform satisfies this condition because its Fourier transform is a rectangular box.

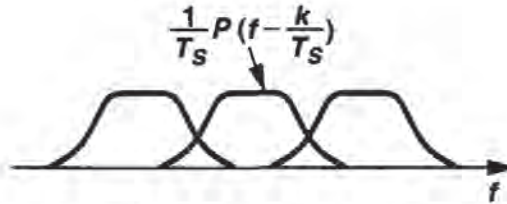


Figure 2.15 Nyquist's condition for the spectrum of a pulse shape that gives no ISI.

A sinc pulse shape, however, introduces difficulties in the design of the system. The filter required to produce the rectangular spectrum becomes quite complex if a sharp cutoff is necessary. Furthermore, the substantial signal energy near the edge of the spectrum complicates the filtering requirements in both the transmit and receive paths. In addition, the sinc waveform decays slowly with time, introducing considerable ISI in the presence of timing errors in the sampling command.

A pulse shape often employed in Nyquist signaling is related to a “raised cosine” spectrum. Shown in Fig. 2.16, the time- and frequency-domain expressions of this function are, respectively,

$$p(t) = \frac{\sin(\pi t/T_s)}{\pi t/T_s} \frac{\cos(\pi \alpha t/T_s)}{1 - 4\alpha^2 t^2/T_s^2} \quad (2.53)$$

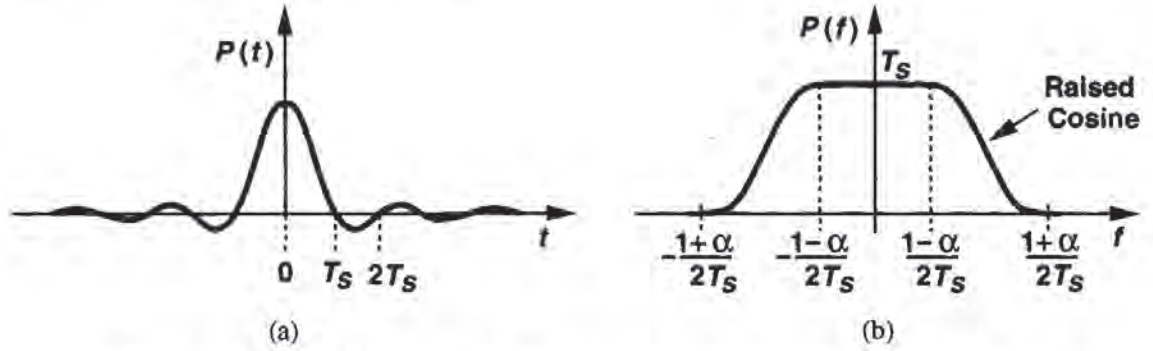


Figure 2.16 Raised-cosine pulse: (a) in time domain, (b) in frequency domain.

and

$$\begin{aligned}
 P(f) &= T_S \quad 0 < |f| < \frac{1-\alpha}{2T_S} \\
 &= \frac{T_S}{2} \left[1 + \cos \frac{\pi T_S}{\alpha} \left(|f| - \frac{1-\alpha}{2T_S} \right) \right] \quad \frac{1-\alpha}{2T_S} < |f| < \frac{1+\alpha}{2T_S} \\
 &= 0 \quad |f| > \frac{1+\alpha}{2T_S},
 \end{aligned} \tag{2.54}$$

where $0 < \alpha < 1$ is the “rolloff” factor. It is interesting to note that (1) $p(t)$ decays faster than a sinc function, (2) for $\alpha = 0$, $p(t)$ reduces to a sinc function, and (3) $P(f)$ is similar to a box spectrum but with smooth edges.

The trade-off in the choice of α is between the decay rate in the time domain and the excess bandwidth (with respect to a box spectrum) in the frequency domain. Typical values of α are between 0.3 and 0.5.

Raised-cosine signaling can also be visualized as shown in Fig. 2.17. Here, each bit is represented by an impulse, and the data stream is applied to a filter whose transfer function is given by (2.54). From this point of view, the operation is called “raised-cosine filtering.”

As explained in Chapter 3, many applications incorporate a filter whose transfer function is equal to the *square root* of that in (2.54).

2.3 RANDOM PROCESSES AND NOISE

Random processes are an integral part of communications, used to represent both signals and noise. In this section, we provide a brief review of random processes and noise to the extent required for and in a language suited to RF design. The primary goal is to develop an intuitive understanding of these phenomena and the relationships governing their behavior. The reader is assumed to be familiar with concepts such as random variables and probability density functions (PDFs).

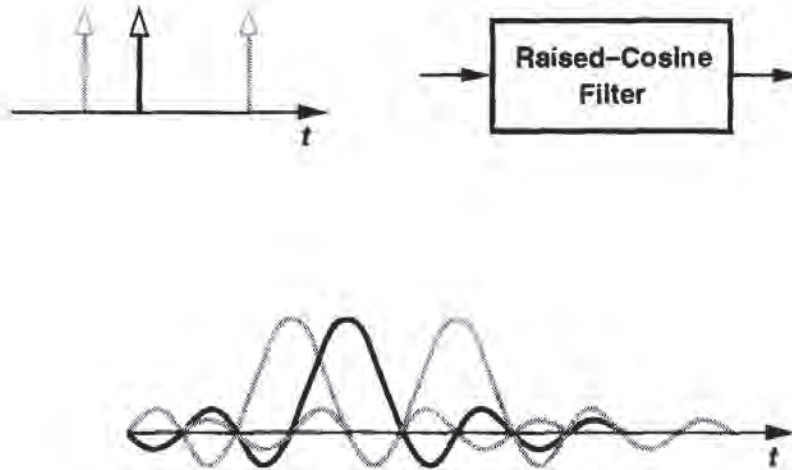


Figure 2.17 Raised-cosine filtering.

2.3.1 Random Processes

The trouble with random processes is that they are random! Engineers who are used to dealing with well-defined, deterministic, “hard facts” often find the concept of randomness difficult to grasp, especially if it must be incorporated mathematically. To overcome this fear of randomness, we approach the problem from an intuitive angle.

We consider a phenomenon random because we do not know or simply do not *need* to know everything about it. We characterize the process with only a few parameters and functions and solve most problems without any other information about the process. Experience shows that this approach is feasible and adequate in many applications, including RF design. In other words, we are fortunate that most random processes encountered in RF design lend themselves to relatively simple modeling.

For our purposes, a random (actually a “stochastic”) process can be defined as “a family of time functions.” If we measure the noise voltage across a resistor as a function of time today, the waveform is different from that measured tomorrow² (Fig. 2.18). To know everything about the noise voltage, we would need to perform an infinite number of measurements, each one for an infinite length of time. Since a single waveform measurement in general does not provide adequate knowledge of the process, even simplest random processes extend in two dimensions; i.e., they require a *collection* of measurements, hence the phrase “family of time functions.” This is the principal difference between random and deterministic signals—and the primary source of confusion. In using an ordinary signal generator, we always consider the output a single predictable and well-defined waveform (except perhaps for the phase at the power-up time, which is usually unimportant). With a random signal, e.g., the

² This should not be confused with time variance in a system.

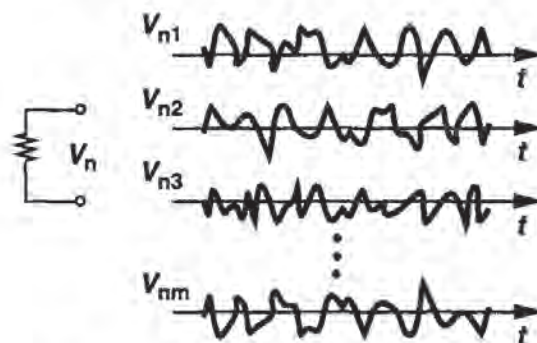


Figure 2.18 Noise viewed as a family of time functions.

voice going through a phone line, we do not have this luxury and must resort to statistics obtained from multiple measurements.

We must emphasize that the reason why the theory of random signals is useful and practical is that in most cases of interest such signals can be modeled with simple statistical functions that indicate, among other things, *how much* and *how fast* the amplitude varies with time. Furthermore, the statistical models can be used in conjunction with the familiar theory of deterministic signals and systems, often allowing us to momentarily forget about randomness and utilize more intuitive analysis techniques.

How is a random process characterized? What aspects of its statistics are important? How are these aspects incorporated in system analysis? We answer these questions by first making some simplifying assumptions.

Statistical Ensembles As mentioned above, full characterization of a (continuous-time) random signal, e.g., the noise voltage across a resistor, requires a “doubly infinite” set: an infinite number of measurements, each for an infinite length of time (Fig. 2.18). Now suppose, rather than one resistor, we consider a very large number of identical resistors and measure their noise voltages simultaneously (still for a very long time). We would expect these two experiments to yield the same statistical results. The large set of resistor noise voltages is called an “ensemble,” and each of the waveforms is called a “sample function.”

How do we measure the average value of the noise voltage of a resistor? Our familiar approach is to measure the noise, $n(t)$, for a long time, T , and calculate the average (or dc component) as

$$\langle n(t) \rangle = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{+T/2} n(t) dt. \quad (2.55)$$

This notion of dc component of a random signal is called the “time average.”

Another definition of the average value is based on simultaneous sampling of all the waveforms in an ensemble (Fig. 2.19). Here, we compute the average by adding the sampled values and normalizing the sum to the number

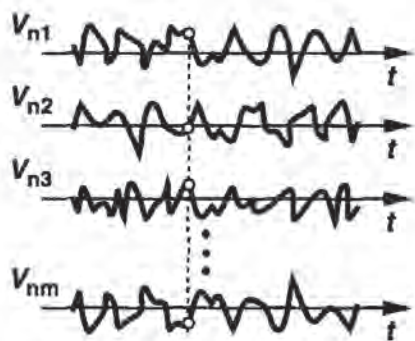


Figure 2.19 Averaging over sample functions.

of waveforms. Called the “ensemble average” or “statistical average,” this type is formally defined as

$$\overline{n(t)} = \int_{-\infty}^{+\infty} n(t) P_n(n) dn, \quad (2.56)$$

where $P_n(n)$ is the probability density function of the process.

From the above definitions arise two questions. First, is the time average measured today equal to that measured tomorrow? Not necessarily. A process whose statistical properties are invariant to a time shift is called “stationary” (more accurately “strict-sense stationary”). Thus, the concept of time average is useful for stationary processes, e.g., noise voltage of a resistor held at a constant temperature. Fortunately, most of the random phenomena in RF systems can be considered stationary.

The second question is: Is the time average of a stationary process equal to the ensemble average? Not always, but for most random processes of interest in this book, we can assume so, thus avoiding ensemble averages.

The time and ensemble averages defined above are of first order. Higher-order averages can also be defined. Of particular interest are second-order averages, for they represent the power of signals. In the time domain,

$$\langle n^2(t) \rangle = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{+T/2} n^2(t) dt, \quad (2.57)$$

which is also called the “mean square” power (with respect to a 1- Ω resistor) if $n(t)$ is a voltage quantity. The second-order ensemble average is

$$\overline{n^2(t)} = \int_{-\infty}^{+\infty} n^2(t) P_n(n) dn. \quad (2.58)$$

For our purposes, $\langle n^2(t) \rangle = \overline{n^2(t)}$.

Probability Density Function When considering a random signal in the time domain, we usually need to know how often its amplitude is between certain limits. For example, if a binary data sequence is corrupted by additive noise (Fig. 2.20), it is important to find the probability that a logical ONE is

interpreted as a ZERO and vice versa, that is, how often the noise amplitude exceeds half of the signal amplitude. The amplitude statistics of a random signal $x(t)$ is characterized by the probability density function, $P_x(x)$, defined as

$$P_x(x)dx = \text{probability of } x < X < x + dx, \quad (2.59)$$

where X is the measured value of $x(t)$ at some point in time. To estimate the PDF, we sample $x(t)$ at many points (for many functions in the ensemble), construct bins of small width, choose the bin height equal to the number of samples whose value falls between the two edges of the bin, and normalize the bin heights to the total number of samples. Note that the PDF provides no information as to *how fast* the random signal varies in the time domain.



Figure 2.20 Binary signal corrupted by noise.

An important example of PDFs is the Gaussian (or normal) distribution. The central limit theorem states that if many independent random processes with arbitrary PDFs are added, the PDF of the sum approaches a Gaussian distribution. It is therefore not surprising that many natural phenomena exhibit Gaussian statistics. For example, since the noise of a resistor results from random “walk” of a very large number of electrons, each having relatively independent statistics, the overall amplitude follows a Gaussian PDF.

The Gaussian PDF is defined as

$$P_x(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(x - m)^2}{2\sigma^2}, \quad (2.60)$$

where σ and m are the standard deviation and mean of the distribution, respectively.

From the PDF of the amplitude of a random signal, we can also answer the following question: If a large number of samples are taken, what percentage will fall between x_1 and x_2 ? This is given by the area under $P_x(x)$ from x_1 to x_2 , and for a Gaussian PDF:

$$P(x_1 < x < x_2) = \int_{x_1}^{x_2} \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(x - m)^2}{2\sigma^2} dx. \quad (2.61)$$

For finite x_1 and x_2 , the integral on the right-hand side must be calculated numerically. A simpler version of this integral, called the error function, is tabulated in many references:

$$\text{erf}(x) = \frac{1}{\sqrt{2\pi}} \int_0^x \exp \frac{-u^2}{2} du. \quad (2.62)$$

It is useful to remember that for a Gaussian distribution approximately 68% of the sampled values fall between $m - \sigma$ and $m + \sigma$ and 99% between $m - 3\sigma$ and $m + 3\sigma$.

Power Spectral Density Since our knowledge of random signals in the time domain is usually quite limited, it is often necessary to characterize such signals in the frequency domain as well. In fact, as we will see throughout this book, the frequency-domain behavior of random signals and noise proves much more useful in RF design than do their time-domain characteristics.

For a deterministic signal $x(t)$, the frequency information is embodied in the Fourier transform:

$$X(f) = \int_{-\infty}^{+\infty} x(t) \exp(-j2\pi ft) dt. \quad (2.63)$$

While it may seem natural to use the same definition for random signals, we must note that the Fourier transform exists only for signals with finite energy,³

$$E_x = \int_{-\infty}^{+\infty} |x(t)|^2 dt < \infty, \quad (2.64)$$

i.e., only if $|x(t)|^2$ drops rapidly enough as $t \rightarrow \infty$. As shown in Fig. 2.21, this condition is violated by two classes of signals: periodic waveforms and random signals. In most cases, however, these waveforms have a finite power:

$$P = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{+T/2} |x(t)|^2 dt < \infty. \quad (2.65)$$

For periodic signals with $P < \infty$, the Fourier transform can still be defined by representing each component of the Fourier series with an impulse in the frequency domain. For random signals, on the other hand, this is generally not possible because a frequency impulse indicates the existence of a deterministic sinusoidal component. Another practical problem is that even if we somehow define a Fourier transform for a random (stationary or nonstationary) process, the result itself is also a random process [5].

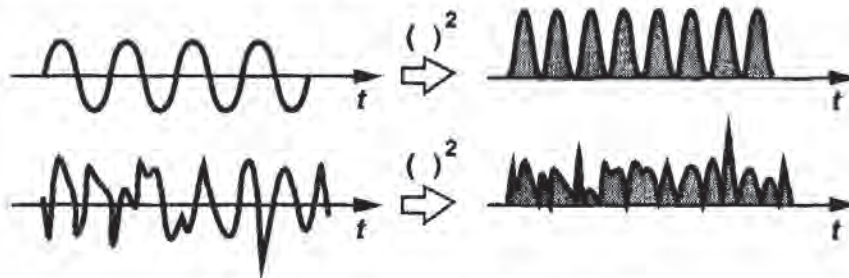


Figure 2.21 Signals with infinite energy.

³ The definition of energy can be visualized if $x(t)$ is a voltage applied across a 1- Ω resistor.

From the above discussion we infer that frequency-domain characteristics of random signals are embodied in a function different from a direct Fourier transform. The power spectral density (PSD) (also called the “spectral density” or simply the “spectrum”) is such a function. Before giving a formal definition of PSD, we describe its meaning from an intuitive point of view [6]. The spectral density, $S_x(f)$, of a random signal $x(t)$ shows how much power the signal carries in a unit bandwidth around frequency f . As illustrated in Fig. 2.22, if we apply the signal to a bandpass filter with a 1-Hz bandwidth centered at f and measure the average output power over a sufficiently long time (on the order of 1 s), we obtain an estimate of $S_x(f)$. If this measurement is performed for each value of f , the overall spectrum of the signal is obtained. This is in fact the principle of operation of spectrum analyzers.⁴

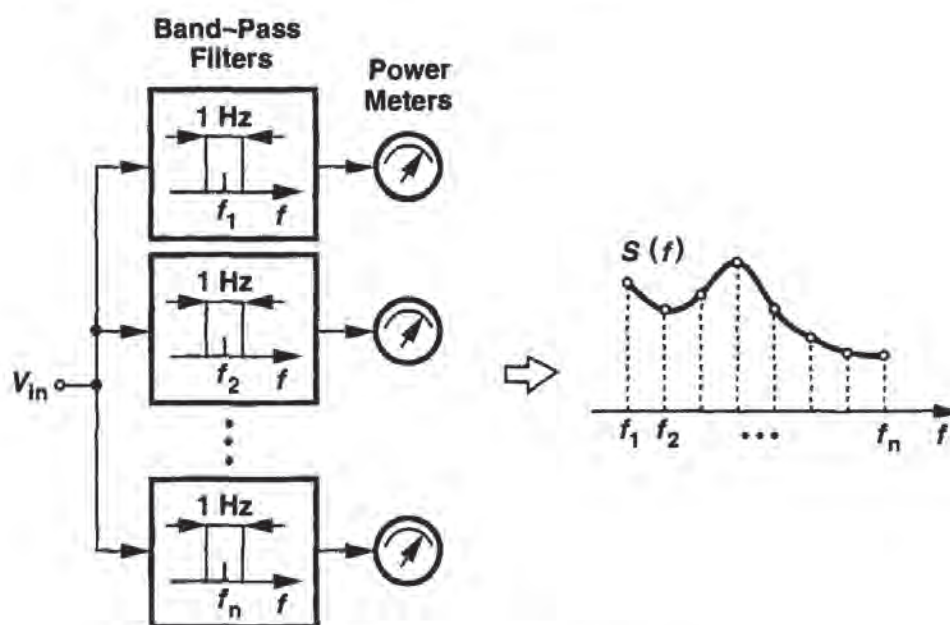


Figure 2.22 Measurement of spectrum.

The formal definition of the PSD is as follows [3]:

$$S_x(f) = \lim_{T \rightarrow \infty} \frac{\overline{|X_T(f)|^2}}{T}, \quad (2.66)$$

where

$$X_T(f) = \int_0^T x(t) \exp(-j2\pi ft) dt. \quad (2.67)$$

⁴ Building a low-loss BPF with 1-Hz bandwidth and a center frequency of, say, 1 GHz is impractical. Thus, actual spectrum analyzers both translate the spectrum to a lower center frequency and measure the power in a band wider than 1 Hz.

The definition can be understood with the aid of a corresponding computational algorithm (Fig. 2.23): (1) truncate $x(t)$ to a relatively long interval $[0, T]$, (2) calculate the Fourier transform of the result and hence $|X_T(f)|^2$, (3) repeat steps 1 and 2 for many sample functions of $x(t)$ (e.g., for many noise voltage waveforms measured across a resistor), and (4) take the average of all $|X_T(f)|^2$ functions to arrive at $\overline{|X_T(f)|^2}$ and normalize the result to T . This algorithm proves useful in time-domain simulations incorporating random noise waveforms.

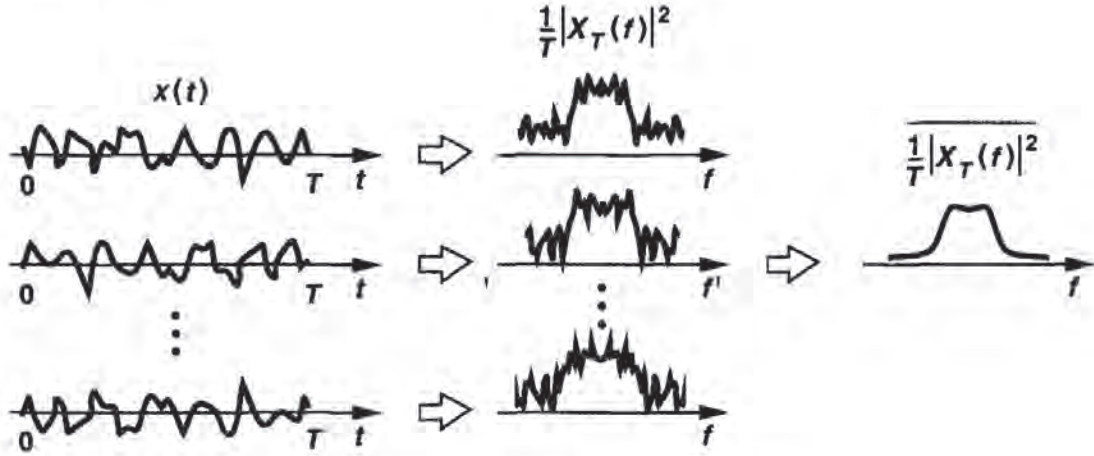


Figure 2.23 Algorithm for PSD estimation.

Since $S_x(f)$ is an even function of f for real $x(t)$ [3], as depicted in Fig. 2.24(a) the total power carried by $x(t)$ in the frequency range $[f_1, f_2]$ is equal to

$$\int_{-f_2}^{-f_1} S_x(f) df + \int_{f_1}^{f_2} S_x(f) df = \int_{f_1}^{f_2} 2S_x(f) df. \quad (2.68)$$

In fact, the right-hand side integral is the quantity measured by a spectrum analyzer; i.e., the negative-frequency part of the spectrum is folded around the vertical axis and is added to the positive-frequency part [Fig. 2.24(b)]. We call the representation of Fig. 2.24(a) the “two-sided” spectrum and that of Fig. 2.24(b) the “one-sided” spectrum.

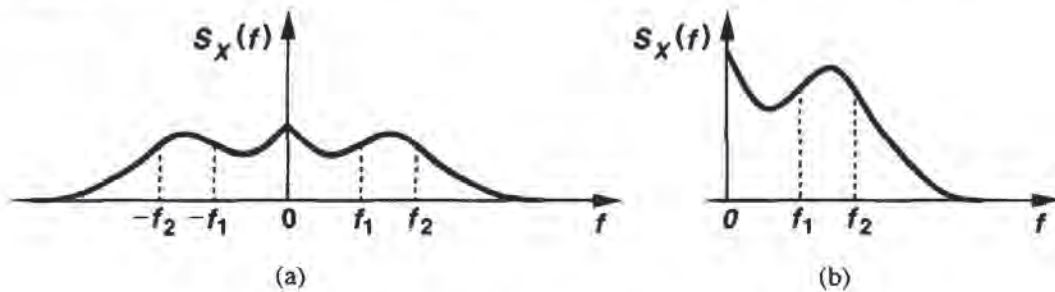


Figure 2.24 (a) Two-sided and (b) one-sided spectra.

In graphical analysis of frequency-domain operations, it is generally simpler to use a two-sided spectrum, whereas actual noise calculations are more easily carried out with a one-sided spectrum. Nevertheless, these two representations bear no fundamental difference—though they can cause confusion.

As an example of $S_x(f)$, we consider the thermal noise voltage across a resistor of value R . The two-sided PSD is

$$S_x(f) = 2kTR, \quad (2.69)$$

where k is the Boltzmann constant and equal to 1.38×10^{-23} J/K and T is the absolute temperature. Such a flat spectrum is called “white” because it contains the same level of power at all frequencies.

Equation (2.69) raises two interesting questions. First, is the total noise power of a resistor [the area under $S_x(f)$] infinite? In reality, $S_x(f)$ is flat for only $|f| < 100$ GHz, dropping beyond this frequency such that the total power remains finite [3]. Second, is the dimension of $2kTR$ power per unit bandwidth (W/Hz)? No, the actual dimension is mean square voltage per unit bandwidth. We tacitly assume that this voltage is applied across a $1\text{-}\Omega$ resistor to generate a power of $2kTR$ in a 1-Hz bandwidth. In circuit noise calculations, we often write

$$\overline{V_n^2} = 4kTR \cdot \Delta f, \quad (2.70)$$

where $\overline{V_n^2}$ is the mean square noise voltage generated by resistor R in a bandwidth Δf . Called the “spot noise” for $\Delta f = 1$ Hz, $\overline{V_n^2}$ is measured in V^2/Hz .

To summarize the concepts of PDF and PSD, we note that the former is a statistical indication of how often the amplitude of a random process falls in a given range of values while the latter shows how much power the signal is expected to contain in a small frequency interval. In general, the PDF and PSD bear no relationship: thermal noise has a Gaussian PDF and a white PSD, whereas flicker ($1/f$) noise has the same type of PDF but a PSD proportional to $1/f$.

Random Signals in Linear Systems The principal reason for defining the power spectral density function is that it allows many of the frequency-domain operations used with deterministic signals to be applied to random processes as well. It can be shown that if a signal with spectral density $S_x(f)$ is applied to a linear time-invariant system with transfer function $H(s)$ (Fig. 2.25), then the output spectrum is

$$S_y(f) = S_x(f)|H(f)|^2, \quad (2.71)$$

where $H(f) = H(s = j2\pi f)$ [3]. This agrees with our intuition that the spectrum of the signal is shaped by the transfer function of the system. It can also be shown that if $x(t)$ is Gaussian, so is $y(t)$ [3].

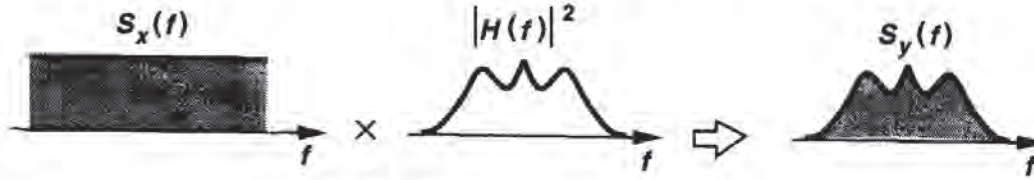


Figure 2.25 Noise shaping in a linear system.

2.3.2 Noise

Noise can be loosely defined as any random interference unrelated to the signal of interest.⁵ This definition distinguishes between noise and deterministic phenomena such as harmonic distortion and intermodulation. As other random processes, noise is characterized by a PDF and a PSD.

Present in all circuits is thermal noise, generated by resistors, base and emitter resistance of bipolar devices, and channel resistance of MOSFETs [Fig. 2.26(a)]. The thermal noise of MOS devices is modeled as a current source connected between the drain and source with a PSD,

$$\overline{I_n^2} = 4kT \left(\frac{2}{3} g_m \right), \quad (2.72)$$

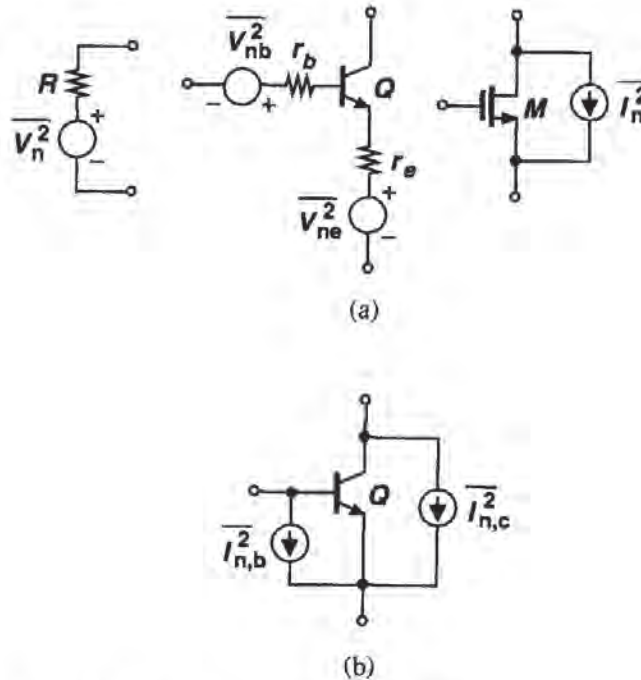


Figure 2.26 (a) Thermal and (b) shot noise in devices.

⁵ It is often said that if there were no noise, there would be no analog designers.

where g_m is the transconductance of the transistor. Derived for long-channel devices [7], the factor $2/3$ may need to be replaced with higher values for channel lengths below $1\ \mu\text{m}$ [8]. Note that the distributed gate resistance of MOSFETs also contributes thermal noise [9], but the effect can be minimized through careful layout.

In addition to thermal noise, active devices may exhibit shot and flicker noise as well. Shot noise is a Gaussian white process associated with the transfer of charge across an energy barrier (e.g., a pn junction) having a PSD

$$\overline{I_n^2} = 2qI, \quad (2.73)$$

where q is the charge of an electron and I the average current. For a bipolar transistor, the collector and base current shot noise is modeled as a current source connected between the collector and emitter and another between the base and the emitter [Fig. 2.26(b)].

Flicker noise arises from random trapping of charge at the oxide-silicon interface of MOSFETs. Represented as a voltage source in series with the gate, the noise density is given by

$$\overline{V_n^2} = \frac{K}{WLC_{ox}} \frac{1}{f}, \quad (2.74)$$

where K is a process-dependent constant. While the effect of flicker noise may seem negligible at high frequencies, we must note that nonlinearity or time variance in circuits such as mixers or oscillators can translate the $1/f$ -shaped spectrum to the RF range (Chapters 6 and 7).

Input-Referred Noise The noise of a two-port system can be modeled by two input noise generators: a series voltage source and a parallel current source (Fig. 2.27) [10]. In general, the correlation between the two sources must be taken into account. We use an example to illustrate the idea. Consider the circuit shown in Fig. 2.28(a), where we assume proper biasing ensures that M_1 is in saturation and carries a drain current of I_D . This circuit has only one dominant source of thermal noise: that due to the channel and represented by $\overline{I_{nD}^2}$. For the model of Fig. 2.28(b), we calculate $\overline{V_n^2}$ by shorting the input port and $\overline{I_n^2}$ by leaving it open. Since the circuits of Figs. 2.28(a) and (b) must produce

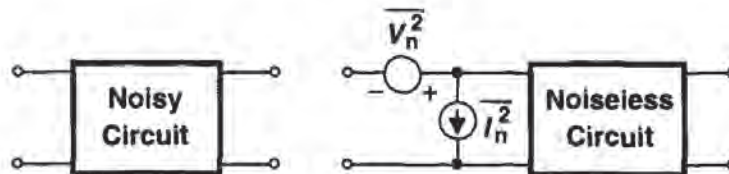


Figure 2.27 Representation of noise by input noise generators.

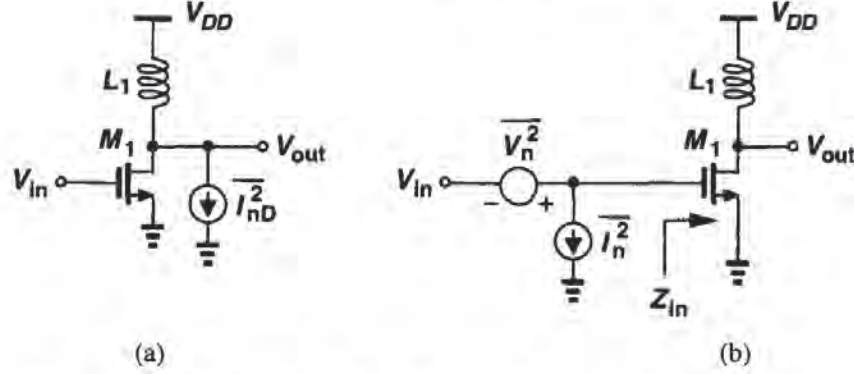


Figure 2.28 (a) MOS amplifier, (b) equivalent input noise generators.

the same output noise in both cases, we have $g_m^2 \overline{V_n^2} = \overline{I_{nD}^2}$ and $g_m^2 \overline{I_n^2} |Z_{in}|^2 = \overline{I_{nD}^2}$, where Z_{in} denotes the input impedance of the circuit. Thus, for $\overline{I_{nD}^2} = 4kT(2g_m/3)$, we obtain $\overline{V_n^2} = 8kT/(3g_m)$ and $\overline{I_n^2} = 8kT/(3g_m|Z_{in}|^2)$. Since V_n and I_n represent the same noise mechanism, they are correlated.

We note that if $|Z_{in}| \rightarrow \infty$, $\overline{I_n^2} \rightarrow 0$, and $\overline{V_n^2}$ is sufficient to represent the noise. At radio frequencies, however, $|Z_{in}|$ is relatively low (in some cases around 50Ω by design), thereby necessitating the use of both $\overline{V_n^2}$ and $\overline{I_n^2}$.

The key point in the above example is that even though the actual circuit may have no physical input noise current, the representation using input-referred sources must include $\overline{I_n^2}$.

Noise Figure In many analog circuits, the signal-to-noise ratio (SNR), defined as the ratio of the signal power to the total noise power, is an important parameter. In RF design, on the other hand, even though the ultimate goal is to maximize the SNR for the received and detected signal, most of the front-end receiver blocks are characterized in terms of their “noise figure” rather than the input-referred noise. This is partly for computational convenience and partly from tradition.

Noise figure has been defined in a number of different ways. The most commonly accepted definition is

$$\text{noise figure} = \frac{SNR_{in}}{SNR_{out}}, \quad (2.75)$$

where SNR_{in} and SNR_{out} are the signal-to-noise ratios measured at the input and output, respectively. Note that the above ratio is called the “noise factor” in most textbooks, with the term noise figure applied to $10 \log_{10}(\text{noise factor})$. We do not make this distinction here.

It is important to understand the physical meaning of (2.75). Noise figure is a measure of how much the SNR degrades as the signal passes through a system. If a system has no noise, then $SNR_{out} = SNR_{in}$, regardless of the gain. This is because both the input signal and the input noise are amplified (or

attenuated) by the same factor and no additional noise is introduced. Therefore, the noise figure of a noiseless system is equal to unity. In reality, the finite noise of a system degrades the SNR, yielding $NF > 1$.

Compared to input-referred noise, the definition of NF in (2.75) may seem rather complicated: it depends on not only the noise of the circuit under consideration but the SNR of the *preceding* stage. In fact, if the input signal contains no noise, $SNR_{in} = \infty$ and $NF = \infty$, even though the circuit may have only a finite internal noise. For such a case, NF is not a meaningful parameter. In RF design, on the other hand, this does not occur because even the signal in the first stage of a receiver is corrupted by the noise due to the radiation resistance of the antenna.

Calculation of the noise figure is generally simpler than (2.75) may suggest. As depicted in Fig. 2.29, we assume SNR_{in} is the ratio of the input signal power to the noise generated by the source resistance, R_S , and modeled by $\overline{V_{RS}^2}$. If the voltage gain from V_{in} to the input port of the circuit (node P) is equal to α , the SNR measured at this node is

$$SNR_{in} = \frac{\alpha^2 \overline{V_{in}^2}}{\alpha^2 \overline{V_{RS}^2}}. \quad (2.76)$$

For a voltage gain of A_v from P to V_{out} , the SNR measured at the output is equal to

$$SNR_{out} = \frac{\alpha^2 A_v^2 \overline{V_{in}^2}}{[\overline{V_{RS}^2} + (\overline{V_n + I_n R_S})^2] \alpha^2 A_v^2} \quad (2.77)$$

$$= \frac{\overline{V_{in}^2}}{[\overline{V_{RS}^2} + (\overline{V_n + I_n R_S})^2]}, \quad (2.78)$$

where V_n and $I_n R_S$ are added before squaring to account for their correlation. It follows that,

$$NF = \frac{\overline{V_{RS}^2} + (\overline{V_n + I_n R_S})^2}{\overline{V_{RS}^2}} \quad (2.79)$$

$$= 1 + \frac{(\overline{V_n + I_n R_S})^2}{\overline{V_{RS}^2}}. \quad (2.80)$$

The NF is usually specified for a 1-Hz bandwidth at a given frequency. Called the “spot” noise figure to emphasize the very small bandwidth, this quantity can be obtained from (2.80) as

$$NF = 1 + \frac{(\overline{V_n + I_n R_S})^2}{4kTR_S}, \quad (2.81)$$

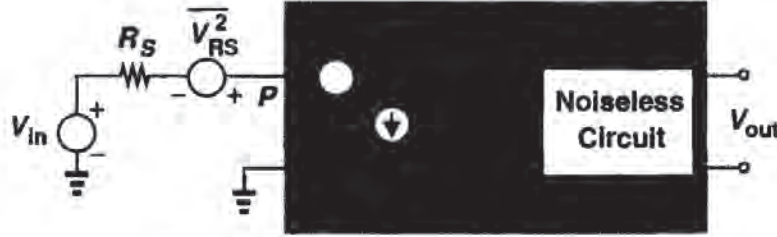


Figure 2.29 Calculation of noise figure.

where V_n and I_n are also measured in unity bandwidth and at the same frequency.

Equation (2.81) indicates that the noise figure is a function of the source impedance, R_S . In general, knowledge of the NF for a given source impedance is not sufficient to calculate the NF for a different source impedance because R_S appears in both the numerator and the denominator of the fraction in (2.81). In traditional RF systems, however, most building blocks are designed so as to exhibit 50- Ω input and output resistance⁶ (with negligible reactance), thereby avoiding ambiguity in NF calculation. As we will see in Chapter 6, this issue still requires attention in certain cases.

For simulation purposes, it is beneficial to write (2.81) as

$$NF = \frac{4kTR_S + (V_n + I_n R_S)^2}{4kTR_S} \quad (2.82)$$

$$= \frac{A^2 [4kTR_S + (V_n + I_n R_S)^2]}{A^2} \frac{1}{4kTR_S} \quad (2.83)$$

$$= \frac{V_{n,out}^2}{A^2} \frac{1}{4kTR_S}, \quad (2.84)$$

where $A = \alpha A_v$ and $V_{n,out}^2$ represents the total noise at the output. Thus, to calculate NF in Fig. 2.29, we divide the total output noise power by the square of the voltage gain from V_{in} to V_{out} and normalize the result to the noise of R_S .

As an example of noise figure calculation, consider the single resistor, R_P , shown in Fig. 2.30(a). What is the noise figure of this circuit with respect to a source resistance R_S ? From Fig. 2.30(b), the total output noise voltage is given by

$$V_{n,out}^2 = 4kT(R_S || R_P), \quad (2.85)$$

and the gain is

$$A = \frac{R_P}{R_S + R_P}. \quad (2.86)$$

⁶ In TV systems, the characteristic and termination impedances are 75 Ω .

It follows from (2.84) that

$$NF = 4kT(R_S || R_P) \frac{(R_S + R_P)^2}{R_P^2} \frac{1}{4kT R_S} \quad (2.87)$$

$$= 1 + \frac{R_S}{R_P}. \quad (2.88)$$

Thus, the noise figure is minimized by maximizing R_P . Interestingly, the condition for minimum noise figure does not coincide with that for maximum power transfer ($R_S = R_P$).

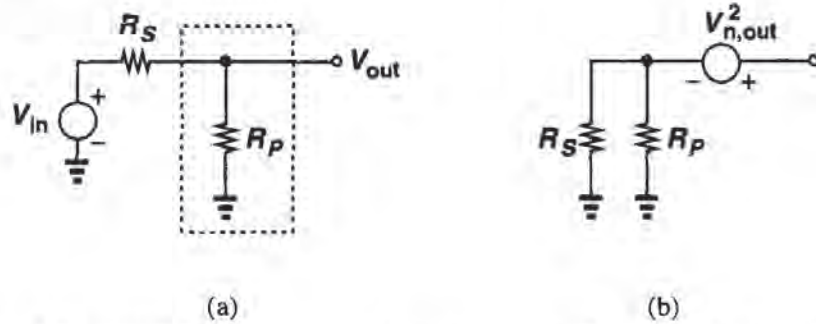


Figure 2.30 (a) Calculation of noise figure of resistor R_P , (b) equivalent circuit of (a).

As another example, let us compute the noise figure of the amplifier shown in Fig. 2.31(a). The circuit consists of a common-source stage and a feedback source follower so as to provide an input resistance equal to R_S . Neglecting body effect, channel length modulation, parasitic capacitances, and the noise of I_1 , we utilize the equivalent circuit shown in Fig. 2.31(b) to calculate the total output noise. The condition $R_{in} = R_S$ translates to

$$R_S = \frac{1}{g_{m2}} \frac{1}{1 + g_{m1} R_D}. \quad (2.89)$$

The noise current of M_2 flows through $R_S/2$, generating an output noise voltage equal to $(I_{n2} R_S/2) g_{m1} R_D$. The noise current of R_D and M_1 is multiplied by the output resistance of the circuit, which for $R_{in} = R_S$ reduces to

$$R_{out} = \frac{R_D}{2} (1 + g_{m2} R_S). \quad (2.90)$$

The total output noise power is therefore equal to

$$\begin{aligned} V_{n,out}^2 &= 4kT R_S \left(\frac{1}{4} g_{m1}^2 R_D^2 \right) + \frac{1}{4} I_{n2}^2 R_S^2 g_{m1}^2 R_D^2 \\ &\quad + (I_{RD}^2 + I_{n1}^2) \frac{R_D^2}{4} (1 + g_{m2} R_S)^2 \end{aligned} \quad (2.91)$$

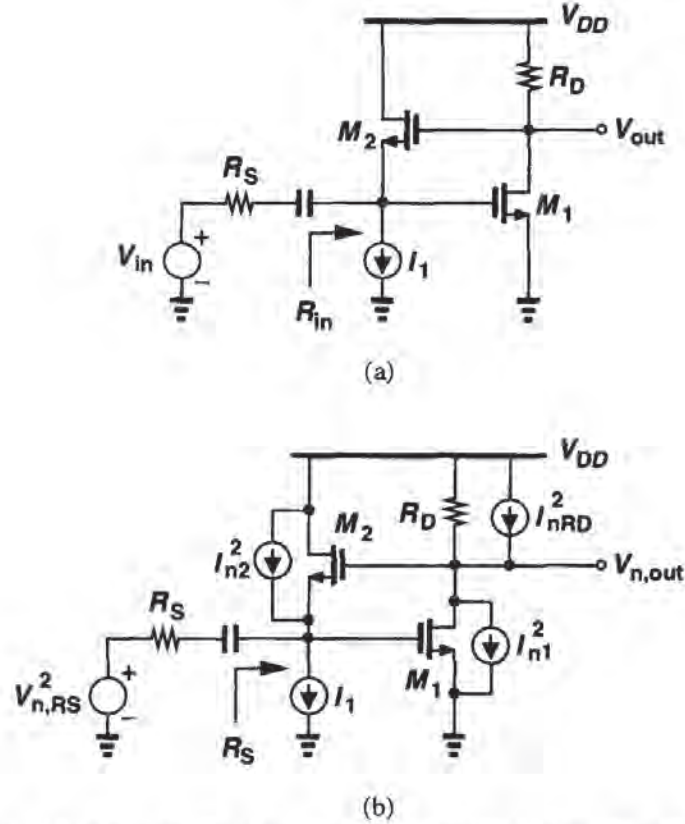


Figure 2.31 (a) Feedback amplifier with input match, (b) noise equivalent circuit of (a).

$$\begin{aligned}
 &= kT R_S g_{m1}^2 R_D^2 + \frac{2}{3} kT g_{m2} R_S^2 g_{m1}^2 R_D^2 \\
 &\quad + \left(\frac{4kT}{R_D} + \frac{8}{3} kT g_{m1} \right) \frac{R_D^2}{4} (1 + g_{m2} R_S)^2. \quad (2.92)
 \end{aligned}$$

Thus, the noise figure is

$$NF = \frac{V_{n,out}^2}{A^2} \frac{1}{4kT R_S} \quad (2.93)$$

$$= 1 + \frac{2}{3} g_{m2} R_S + \left(\frac{1}{R_D} + \frac{2}{3} g_{m1} \right) \frac{(1 + g_{m2} R_S)^2}{g_{m1}^2 R_S}, \quad (2.94)$$

subject to the condition $g_{m2} R_S = (1 + g_{m1} R_D)^{-1}$.

Noise Figure of Cascaded Stages For a cascade of stages, the overall noise figure can be obtained in terms of the NF and gain of each stage. Consider the system shown in Fig. 2.32, where input noise generators and input and output resistances of each stage are shown. Note that reactive components of the impedances are nulled and A_{v1} and A_{v2} denote the *unloaded* voltage gain

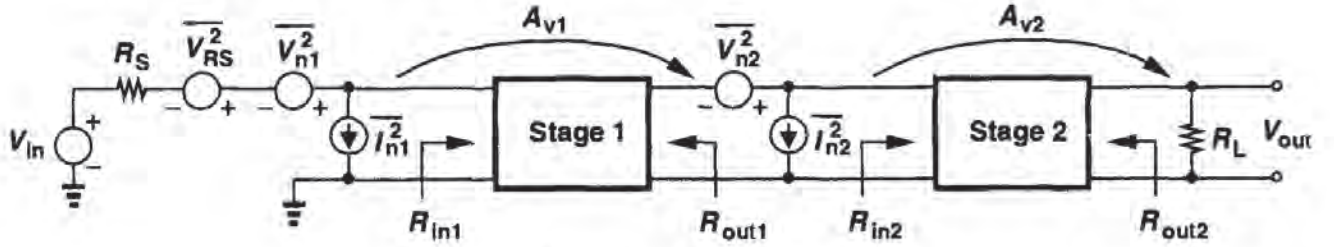


Figure 2.32 Cascaded noisy stages.

of the two stages. The total noise power at the input of the first stage can be written as

$$V_{n,in1}^2 = [I_{n1}(R_S || R_{in1}) + V_{n1} \frac{R_{in1}}{R_{in1} + R_S}]^2 + \overline{V_{RS}^2} \frac{R_{in1}^2}{(R_{in1} + R_S)^2}. \quad (2.95)$$

We also note that the total noise power at the input of the second stage is

$$V_{n,in2}^2 = V_{n,in1}^2 A_{v1}^2 \left(\frac{R_{in2}}{R_{out1} + R_{in2}} \right)^2 + \left[I_{n2}(R_{out1} || R_{in2}) + V_{n2} \frac{R_{in2}}{R_{in2} + R_{out1}} \right]^2. \quad (2.96)$$

Thus, the total output noise power of the cascade equals $A_{v2}^2 V_{n,in2}^2 R_L^2 / (R_L + R_{out2})^2$. Since the total voltage gain from V_{in} to V_{out} equals

$$A_{v,tot} = \frac{R_{in1}}{R_S + R_{in1}} A_{v1} \frac{R_{in2}}{R_{out1} + R_{in2}} A_{v2} \frac{R_L}{R_{out2} + R_L}, \quad (2.97)$$

the overall noise figure is

$$NF_{tot} = \frac{1}{A_{v,tot}^2} A_{v2}^2 \left(\frac{R_L}{R_L + R_{out2}} \right)^2 V_{n,in2}^2 \cdot \frac{1}{4kTR_S}. \quad (2.98)$$

Using (2.95) and (2.96) and simplifying the result, we have

$$NF_{tot} = \frac{4kTR_S + (I_{n1}R_S + V_{n1})^2}{4kTR_S} + \frac{(I_{n2}R_{out1} + V_{n2})^2}{A_{v1}^2} \frac{1}{\left(\frac{R_{in1}}{R_S + R_{in1}} \right)^2} \frac{1}{4kTR_S}. \quad (2.99)$$

The first term on the right-hand side can be identified as the NF of the first stage with respect to a source impedance R_S . The second term, on the other hand,

is not as straightforward. In the special case where $R_S = R_{in1} = R_{out1} = R_{in2}$, we have

$$NF_{tot} = NF_1 + \frac{(I_{n2}R_S + V_{n2})^2}{\alpha^2 A_{v1}^2} \frac{1}{4kTR_S} \quad (2.100)$$

$$= NF_1 + \frac{NF_2 - 1}{\alpha^2 A_{v1}^2}, \quad (2.101)$$

where NF_2 is the noise figure of the second stage with respect to a source impedance R_S .

In the general case, we simplify (2.99) using the concept of “available power gain,” A_P . This type of gain is defined as the available power at the output (the power that the circuit would deliver to a conjugate-matched load) divided by the available source power (the power that the source would deliver to a conjugate-matched circuit.) The available output power of stage 1 in Fig. 2.32 is

$$P_{out,av} = V_{in}^2 \left(\frac{R_{in1}}{R_S + R_{in1}} \right)^2 A_{v1}^2 \cdot \frac{1}{4R_{out1}}, \quad (2.102)$$

and the available source power is

$$P_{source,av} = \frac{V_{in}^2}{4R_S}. \quad (2.103)$$

Thus,

$$A_P = \left(\frac{R_{in1}}{R_S + R_{in1}} \right)^2 A_{v1}^2 \frac{R_S}{R_{out1}}. \quad (2.104)$$

Noting that the noise figure of stage 2 with respect to a source impedance R_{out1} is

$$NF_{2,Rout1} = 1 + \frac{(I_{n2}R_{out1} + V_{n2})^2}{4kTR_{out1}}, \quad (2.105)$$

we can write (2.99) as

$$NF_{tot} = NF_{1,RS} + \frac{NF_{2,Rout1} - 1}{A_P}, \quad (2.106)$$

where $NF_{1,RS}$ denotes the noise figure of stage 1 with respect to a source impedance R_S .

Similarly, for m stages,

$$NF_{tot} = 1 + (NF_1 - 1) + \frac{NF_2 - 1}{A_{p1}} + \cdots + \frac{NF_m - 1}{A_{p1} \cdots A_{p(m-1)}}, \quad (2.107)$$

where the NF of each stage is calculated with respect to the source impedance driving that stage. This is called the Friis equation [11]. Expressing the overall

noise figure in terms of the noise figure of each stage, this relation proves especially useful if a receiver employs various off-the-shelf building blocks that are characterized independently by manufacturers.

The Friis equation indicates that the noise contributed by each stage decreases as the gain preceding the stage increases, implying that the first few stages in a cascade are the most critical. Conversely, if a stage exhibits attenuation (loss), then the noise figure of the following circuit is “amplified” when referred to the input of that stage. This occurs, for example, if a narrowband lossy filter is interposed between the antenna and the low noise amplifier in a receiver to reject out-of-band interferers (Chapter 5).

Noise Figure of Lossy Circuits Passive filters used in RF receivers have a finite in-band loss. In addition to attenuating the desired signal, lossy circuits in general contribute noise as well, a fact that may not be obvious if we consider an ideal LC filter [Fig. 2.33(a)] as an example. Recall, however, that many RF circuits are required to have well-defined *resistive* input and output impedances for proper matching. For example, the filter placed between the antenna and the LNA can be viewed as depicted in Fig. 2.33(b). Our goal is to find the relationship between the noise figure and the loss of a passive circuit with resistive input and output impedances.

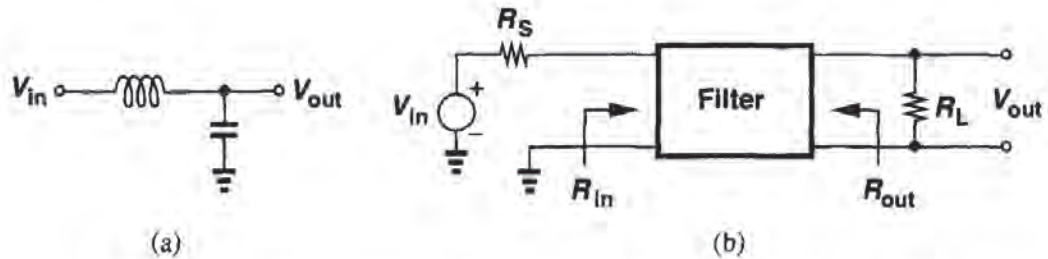


Figure 2.33 (a) LC attenuator, (b) lossy circuit matched at input and output.

Consider a linear time-invariant passive reciprocal network as shown in Fig. 2.34(a) with real input and output impedance. It can be proved [5] that if the output resistance is R_{out} , then the noise Thevenin equivalent circuit is as depicted in Fig. 2.34(b), with the PSD of the voltage source given by $4kTR_{out}$. Note that R_{out} and hence $\overline{V_n^2}$ generally depend on the source impedance, R_S .

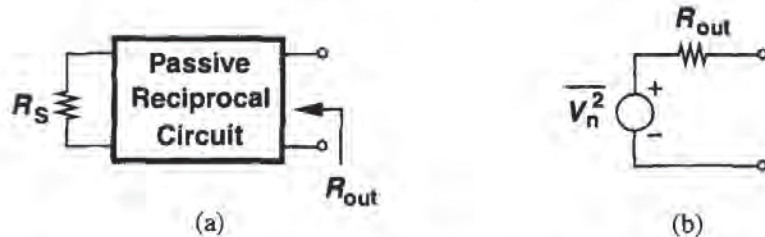


Figure 2.34 (a) Passive reciprocal network, (b) Thevenin noise model of (a).

Now consider the circuit of Fig. 2.35(a). In analogy with the concept of available power gain, we define the power loss L as $P_{\text{in}}/P_{\text{out}}$, where P_{in} is the available source power and P_{out} the available power at the output. To calculate the signal loss, the circuit can be modeled by the Thevenin equivalent shown in Fig. 2.35(b). Note that this circuit is the equivalent for the *signal*, whereas that in Fig. 2.34(b) is for the noise. Since $P_{\text{in}} = V_{\text{in}}^2/(4R_S)$ and $P_{\text{out}} = V_{\text{TH}}^2/(4R_{\text{out}})$, we have

$$L = \frac{V_{\text{in}}^2}{V_{\text{TH}}^2} \frac{R_{\text{out}}}{R_S}. \quad (2.108)$$

To compute the noise figure, we find the output noise voltage and the voltage gain. With a load R_L , Fig. 2.34(b) implies that

$$V_{n,\text{out}}^2 = 4kT R_{\text{out}} \frac{R_L^2}{(R_L + R_{\text{out}})^2}. \quad (2.109)$$

The voltage gain from V_{in} to V_{out} in Fig. 2.35(b) is

$$A_v = \frac{V_{\text{TH}}}{V_{\text{in}}} \frac{R_L}{R_L + R_{\text{out}}}. \quad (2.110)$$

Thus,

$$NF = 4kT R_{\text{out}} \frac{V_{\text{in}}^2}{V_{\text{TH}}^2} \frac{1}{4kT R_S} \quad (2.111)$$

$$= L. \quad (2.112)$$

We conclude that for a passive reciprocal network the noise figure is equal to the loss if the latter is defined as above.

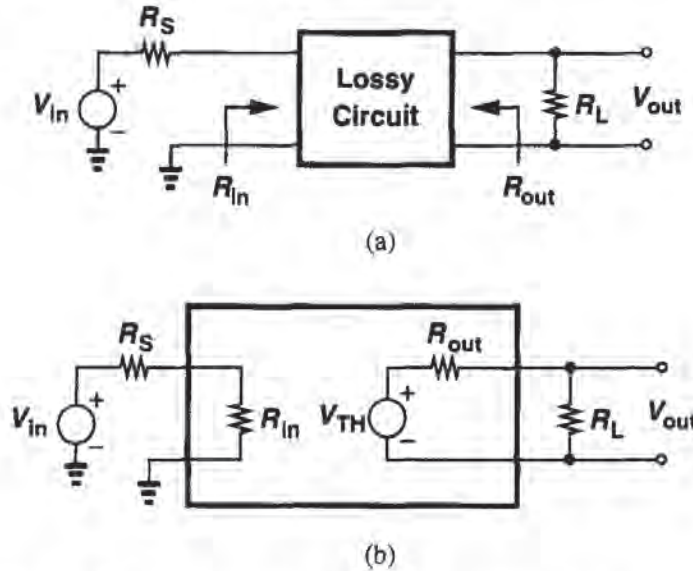


Figure 2.35 (a) Circuit for noise figure calculation, (b) equivalent circuit of (a).

To reinforce the above concepts, we calculate the overall noise figure of a lossy filter followed by a low-noise amplifier (Fig. 2.36). From the Friis equation,

$$NF_{\text{tot}} = NF_{\text{filt}} + \frac{NF_{\text{LNA}} - 1}{L^{-1}} \quad (2.113)$$

$$= L + (NF_{\text{LNA}} - 1)L \quad (2.114)$$

$$= L \cdot NF_{\text{LNA}}, \quad (2.115)$$

where the noise figure of the LNA is calculated with respect to the output resistance of the filter.

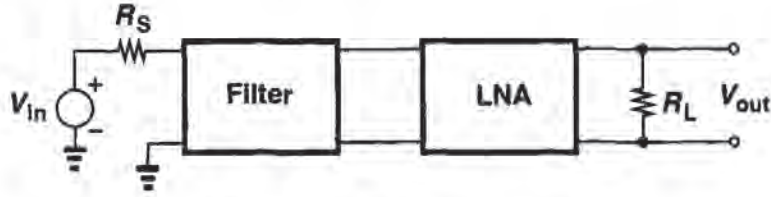


Figure 2.36 Cascade of lossy filter and LNA.

2.4 SENSITIVITY AND DYNAMIC RANGE

Sensitivity The sensitivity of an RF receiver is defined as the minimum signal level that the system can detect with acceptable signal-to-noise ratio. To calculate the sensitivity, we write

$$NF = \frac{SNR_{\text{in}}}{SNR_{\text{out}}} \quad (2.116)$$

$$= \frac{P_{\text{sig}}/P_{RS}}{SNR_{\text{out}}}, \quad (2.117)$$

where P_{sig} denotes the input signal power and P_{RS} the source resistance noise power, both per unit bandwidth. It follows that

$$P_{\text{sig}} = P_{RS} \cdot NF \cdot SNR_{\text{out}}. \quad (2.118)$$

Since the overall signal power is distributed across the channel bandwidth, B , the two sides of (2.118) must be integrated over the bandwidth to obtain the total mean square power. Thus, for a flat channel,

$$P_{\text{sig,tot}} = P_{RS} \cdot NF \cdot SNR_{\text{out}} \cdot B. \quad (2.119)$$

Equation (2.119) predicts the sensitivity as the minimum input signal that yields a given value for the output SNR . Changing the notation slightly and expressing the quantities in dB or dBm, we have

$$P_{\text{in,min}}|_{\text{dBm}} = P_{RS}|_{\text{dBm/Hz}} + NF|_{\text{dB}} + SNR_{\text{min}}|_{\text{dB}} + 10 \log B, \quad (2.120)$$

where $P_{\text{in},\text{min}}$ is the minimum input level that achieves SNR_{min} and B is expressed in hertz. Note that (2.120) does not depend on the gain of the system.

Assuming conjugate matching at the input, we obtain P_{RS} as the noise power that R_S delivers to the receiver:

$$P_{RS} = \frac{4kTR_S}{4} \frac{1}{R_{\text{in}}} \quad (2.121)$$

$$= kT \quad (2.122)$$

$$= -174 \text{ dBm/Hz} \quad (2.123)$$

at room temperature. We thus simplify (2.120) as

$$P_{\text{in},\text{min}} = -174 \text{ dBm/Hz} + NF + 10 \log B + SNR_{\text{min}}. \quad (2.124)$$

Note that the sum of the first three terms is the total integrated noise of the system and is sometimes called the “noise floor.” Since $P_{\text{in},\text{min}}$ is a function of the bandwidth, a receiver may appear very sensitive simply because it employs a narrowband channel (but of course at the cost of low information rate.)

Dynamic Range Dynamic range (DR) is generally defined as the ratio of the maximum input level that the circuit can tolerate to the minimum input level at which the circuit provides a reasonable signal quality. This definition is quantified in different applications differently. For example, in analog circuits such as op amps and analog-to-digital converters the dynamic range is defined as the ratio of the “full-scale” (FS) input level to the the input level for which $SNR = 1$. The full scale is typically the input level beyond which a hard saturation occurs and can be easily found by examining the circuit, and the minimum input level is determined by the noise floor.

In RF design, on the other hand, the situation is more complicated. Consider a simple common-source stage. How do we define the input full scale for such a circuit? It is possible to define the FS as the input voltage for which the transistor is at the edge of triode region. However, if a sinusoid with full-scale swing is applied to the circuit, the output exhibits substantial distortion. Also, the minimum signal must provide an SNR greater than unity, for example SNR_{min} in Eq. (2.120). For these reasons, we base the definition of the upper end of the dynamic range on the intermodulation behavior and the lower end on the sensitivity. Such a definition is called the “spurious-free dynamic range” (SFDR).

The upper end of the dynamic range is defined as the maximum input level in a two-tone test for which the third-order IM products do not exceed the noise floor. Expressing all of the quantities in dBm, we can rewrite (2.33) as

$$P_{IIP3} = P_{\text{in}} + \frac{P_{\text{out}} - P_{IM,\text{out}}}{2}, \quad (2.125)$$

where $P_{IM,out}$ denotes the power of IM_3 components at the output. Since $P_{out} = P_{in} + G$ and $P_{IM,out} = P_{IM,in} + G$, where G is the circuit's power gain in dB and $P_{IM,in}$ is the input-referred level of the IM_3 products, we have

$$P_{IIP3} = P_{in} + \frac{P_{in} - P_{IM,in}}{2} \quad (2.126)$$

$$= \frac{3P_{in} - P_{IM,in}}{2}, \quad (2.127)$$

and hence

$$P_{in} = \frac{2P_{IIP3} + P_{IM,in}}{3}. \quad (2.128)$$

The input level for which the IM products become equal to the noise floor is thus given by

$$P_{in,max} = \frac{2P_{IIP3} + F}{3} \quad (2.129)$$

where $F = -174 \text{ dBm} + NF + 10 \log B$.

The SFDR is the difference (in dB) between $P_{in,max}$ and $P_{in,min}$:

$$SFDR = \frac{2P_{IIP3} + F}{3} - (F + SNR_{min}) \quad (2.130)$$

$$= \frac{2(P_{IIP3} - F)}{3} - SNR_{min}. \quad (2.131)$$

For example, if a receiver with $NF = 9 \text{ dB}$, $P_{IIP3} = -15 \text{ dBm}$, and $B = 200 \text{ kHz}$ requires an $SNR_{min} = 12 \text{ dB}$, then $SFDR \approx 53 \text{ dB}$.

The spurious-free dynamic range represents the maximum relative level of interferers that a receiver can tolerate while producing an acceptable signal quality from a small input level.

2.5 PASSIVE IMPEDANCE TRANSFORMATION

At radio frequencies, we often resort to passive circuits to transform impedances—from high to low and vice versa or from complex to real and vice versa. While active devices may also seem a plausible choice for such operations, in some cases only passive components can achieve the required performance. For example, a 3-V power amplifier can deliver a maximum of $P = 3^2/(2 \times 50) = 90 \text{ mW}$ to a $50\text{-}\Omega$ load. To increase the output power, a circuit must be interposed between the PA and the load so as to “amplify” the voltage swings while requiring no higher supply voltage (Chapter 9). This is performed by passive matching circuits.

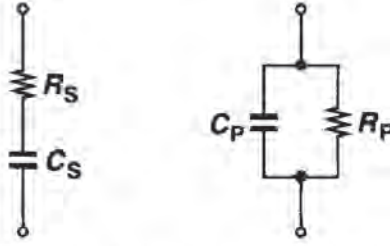


Figure 2.37 Equivalent series and parallel RC circuits.

Before studying transformation techniques, let us consider the RC circuits of Fig. 2.37. The quality factor Q of the series combination, defined as the impedance of the capacitor divided by the resistor, is equal to $1/(R_S C_S \omega)$, approaching infinity as R_S goes to zero. Similarly, the Q of the parallel combination is equal to $R_P C_P \omega$. If Q is relatively high (greater than approximately 5) and the band of interest relatively narrow, then one network can be converted to the other. The two circuits are equivalent if

$$\frac{R_P}{R_P C_P s + 1} = \frac{R_S C_S s + 1}{C_S s}, \quad (2.132)$$

or, for $s = j\omega$, $R_P C_P = 1/(R_S C_S \omega^2)$ and $R_P C_P + R_S C_S - R_P C_S = 0$. Assuming $R_P \gg R_S$, we have $C_P \approx C_S$ and

$$R_P \approx \frac{1}{R_S (C \omega)^2}, \quad (2.133)$$

where $C = C_P \approx C_S$. Thus, the conversion changes the value of the resistance according to (2.133) while keeping the value of the capacitance nearly constant. We can also write $R_P \approx Q_S^2 R_S$, where Q_S is the Q of the series network. Similar results can easily be derived for RL counterparts.

Transformation of impedance can be accomplished by transformers. An ideal transformer with a turn ratio of m scales an impedance by a factor m^2 . In reality, however, high-frequency transformers exhibit loss, capacitive coupling between the primary and the secondary, and even unwanted resonances, thus complicating the design and requiring careful modeling. For this reason, we study other approaches to impedance transformation.

Consider the network shown in Fig. 2.38(a), where the capacitive divider is utilized to transform R_P to a higher value. With the assumptions of high Q and narrow bandwidth, the parallel combination of C_P and R_P can be converted to the series circuit shown in Fig. 2.38(b), where $C_S \approx C_P$ and $R_S \approx 1/[R_P (C_P \omega)^2]$. Combining C_1 and C_S into C_{eq} , we arrive at the circuit of Fig. 2.38(c), which can be converted to the parallel network of Fig. 2.38(d), with $C_{tot} \approx C_1 C_P / (C_1 + C_P)$ and $R_{tot} \approx 1/[R_S (C_{eq} \omega)^2] = (1 + C_P / C_1)^2 R_P$. Thus, the capacitive divider “boosts” the value of R_P by a factor $(1 + C_P / C_1)^2$.

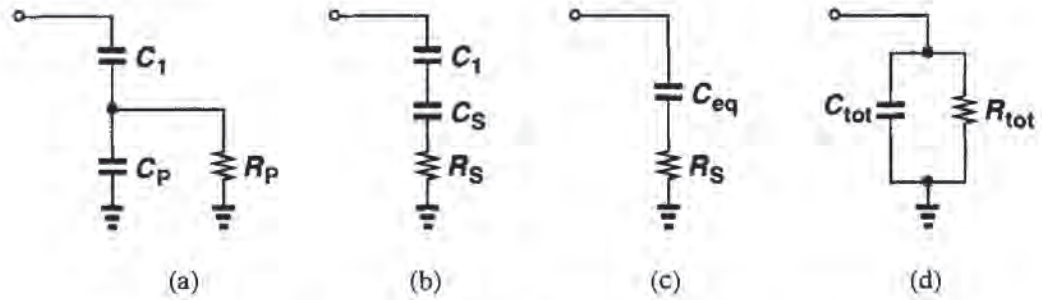


Figure 2.38 Impedance transformation by means of a capacitive divider.

Fig. 2.39(a) depicts a similar transformation circuit using inductive voltage division. For the equivalent circuit of Fig. 2.39(b), we have $L_{\text{tot}} \approx L_1 + L_P$ and $R_{\text{tot}} \approx (1 + L_1/L_P)^2 R_P$ if the Q is high and the band of interest narrow.

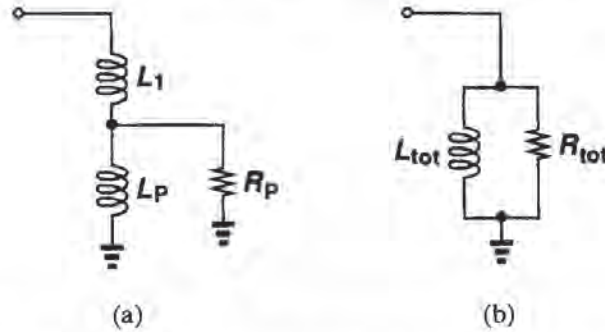


Figure 2.39 Impedance transformation by means of an inductive divider.

A network often employed to transform a resistance to a lower value is illustrated in Fig. 2.40(a). Converting C_P and R_P to the series combination shown in Fig. 2.40(b), we have $C_S \approx C_P$ and $R_S \approx 1/[R_P(C_P\omega)^2]$. In the vicinity of resonance, L_1 and C_S resonate and the network is approximately equivalent to a resistor equal to $1/(C_P^2\omega^2 R_P)$.

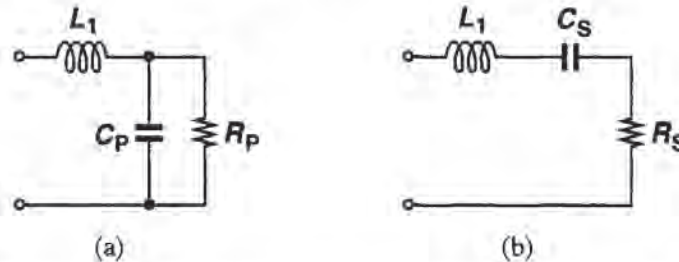


Figure 2.40 Transformation of a resistance to a lower value.

We should note that the assumption of high Q is not always valid, particularly if on-chip inductors are used. Thus, the accuracy of the foregoing derivations must be checked in such cases.

REFERENCES

- [1] D. D. Wiener and J. F. Spine, *Sinusoidal Analysis and Modeling of Weakly Nonlinear Circuits*, New York: Van Nostrand Reinhold, 1980.
- [2] S. A. Maas, *Nonlinear Microwave Circuits*, Norwood, MA: Artech House, 1988.
- [3] L. W. Couch, *Digital and Analog Communication Systems*, 4th ed., New York: Macmillan, 1993.
- [4] J. G. Proakis and M. Salehi, *Communication System Engineering*, Englewood Cliffs, NJ: Prentice Hall, 1994.
- [5] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed., New York: McGraw-Hill, 1991.
- [6] R. W. Bennet, "Methods of Solving Noise Problems," *Proc. IRE*, Vol. 44, pp. 609–638, May 1956.
- [7] A. van der Ziel, "Thermal Noise in Field Effect Transistors," *Proc. IRE*, Vol. 50, pp. 1808–1812, August 1962.
- [8] A. A. Abidi, "High-Frequency Noise Measurements on FETs with Small Dimensions," *IEEE Trans. Electron Devices*, Vol. 33, pp. 1801–1805, November 1986.
- [9] B. Razavi, "Impact of Distributed Gate Resistance on the Performance of MOS Devices," *IEEE Trans. Circuits and Systems—Part I*, Vol. 41, pp. 750–754, November 1994.
- [10] P. R. Gray and R. G. Meyer, *Analysis and Design of Analog Integrated Circuits*, 3rd ed., New York: John Wiley, 1993.
- [11] H. T. Friis, "Noise Figure of Radio Receivers," *Proc. IRE*, Vol. 32, pp. 419–422, July 1944.

MODULATION AND DETECTION

Analog and digital modulation and detection are essential functions in communication systems. While new methods of performing these functions continue to appear, certain schemes have matured over the years and are employed in many RF applications.

This chapter deals with basic modulation and detection techniques used in today's RF transceivers, describing the merits and limitations of each method and emphasizing the implications in the design of RF ICs. Following a brief description of important aspects of modulation schemes, we review analog amplitude, phase, and frequency modulation. Next, we study binary and quadrature digital modulation techniques, focusing on their role in RF communications. Finally, we describe the power efficiency of various modulation schemes and consider noncoherent detection.

3.1 GENERAL CONSIDERATIONS

The transmitted waveform in RF communications is usually a high-frequency carrier modulated by the original signal. Various reasons exist for modulation: (1) in wired systems, coaxial lines exhibit superior shielding at higher frequencies; (2) in wireless systems, the antenna size should be a significant fraction of the wavelength to achieve a reasonable gain; (3) in most cases, the communication must occur in a certain part of the spectrum because of FCC regulations; and (4) in some applications, modulation allows simpler detection at the receive end in the presence of nonidealities in the communication channel.

In the context of modulation, it is common practice to define two types of signals. A "baseband" signal is defined as one whose spectrum is nonzero in the vicinity of $\omega = 0$ and negligible elsewhere [1] [Fig. 3.1(a)], for example, the signal generated by a microphone or a video camera. A "passband"

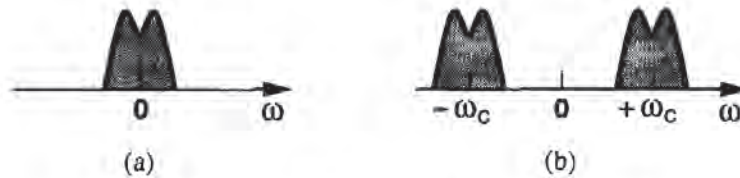


Figure 3.1 (a) Baseband and (b) passband signals.

signal¹ is a waveform whose spectrum is nonzero in a band around a “carrier” frequency ω_c and negligible outside this band [1] [Fig. 3.1(b)]. While mathematically imprecise, these definitions are adequate for our purposes because in RF design the signal bandwidths of interest are usually much less than the carrier frequencies.

Modulation converts a baseband signal to a passband counterpart. From another point of view, modulation varies certain parameters of a carrier (e.g., a sinusoid) according to the baseband signal. These two definitions are equivalent because a passband signal can always be expressed as

$$x(t) = a(t) \cos[\omega_c t + \theta(t)], \quad (3.1)$$

where $a(t)$ and $\theta(t)$ are generally functions of time [1]. In other words, we begin with a periodic carrier signal $A \cos(\omega_c t + \theta)$ and vary its amplitude or phase to perform modulation. In (3.1), the argument $\omega_c t + \theta(t)$ is called the “total phase” and $\theta(t)$ the “excess phase.” We also define the instantaneous frequency as the time derivative of the phase: $\omega_c + d\theta/dt$ is the “total frequency” and $d\theta/dt$ is the “excess frequency” or the “frequency deviation.” Note that, depending on the modulation scheme, a baseband signal with bandwidth B may occupy a different bandwidth when converted to a passband waveform. This will be clarified later.

The inverse of modulation is demodulation or detection, with the goal being to extract the original baseband signal with minimum noise, distortion, ISI, etc. Thus, as depicted in Fig. 3.2, a simple communication system consists of a modulator/transmitter, a channel (e.g., air or coaxial cable), and a

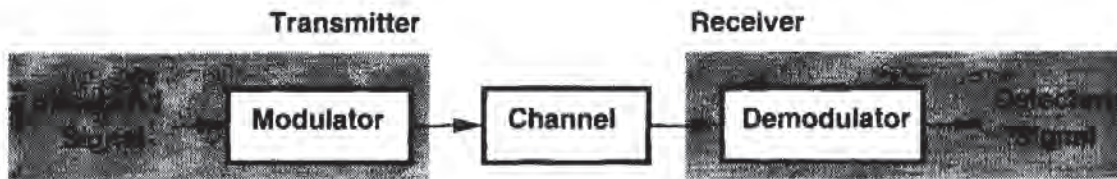


Figure 3.2 Simple communication system.

¹ The term “bandpass” is also used. Passband is consistent with baseband and bandpass with low-pass.

receiver/detector. Since the overall performance depends on the design of both the modulator and the detector, modulation schemes are usually studied in conjunction with proper demodulators. We refer to the combination of the two as “modems.”

Important Aspects of Modems In analyzing the system shown in Fig. 3.2, the primary objective is to find the “quality” (e.g., SNR) of the output of the detector in the presence of attenuation and interference in the channel as well as noise at the input of the detector. For a given channel attenuation (also called “path loss”) and transmitted power, the quality depends on the type of the modem. Thus, if the modem achieves higher tolerance of noise, then either the transmitted power can be reduced, providing longer “talk time” in portable devices, or a higher path loss can be accommodated, allowing communication over a longer distance.

Another important property of modulation schemes is the required bandwidth. Assuming a certain bandwidth for the baseband signal, we calculate that occupied by the modulated waveform in each case. Called the “spectral efficiency,” this criterion plays an important role in choosing a modem for limited-bandwidth applications such as wireless networks.

The third attribute of modulation techniques that is of great concern in RF design is the “power efficiency,” a reference to the type of power amplifier that can be used in the transmitter. As we will see later in this chapter, some modulated waveforms can be processed by means of nonlinear amplifiers without spectral spreading or loss of information whereas some others require linear amplifiers. Since nonlinear power amplifiers are generally more efficient than their linear counterparts (Chapter 9), it is desirable to employ a modulation scheme that lends itself to nonlinear amplification.

In summary, the choice of modems is based on three important aspects: quality, spectral efficiency, and power efficiency.

Additive White Gaussian Noise Channel In order to compare the “noise tolerance” of different receivers, we assume the transmitted signal power is constant and the received signal is corrupted by additive white Gaussian noise (AWGN) having a power spectral density equal to $N_0/2$ (Fig. 3.3). The AWGN assumption proves adequate in taking into account the inherent noise of the receiver. However, the noise introduced by adjacent interferers is generally not white and must be considered separately. Furthermore, the noise in some demodulators (e.g., envelope detectors) is multiplicative rather than additive.

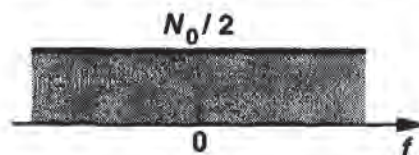


Figure 3.3 White spectrum assumed for additive Gaussian noise.

3.2 ANALOG MODULATION

3.2.1 Amplitude Modulation

For a baseband signal $x_{BB}(t)$, an amplitude-modulated (AM) waveform can be constructed by writing (3.1) as $x_{AM}(t) = A_c[1 + m x_{BB}(t)] \cos \omega_c t$, where m is called the “modulation index.” Illustrated in Fig. 3.4 is a method of generating an AM signal along with the corresponding waveforms and spectra: multiplication by $\cos \omega_c t$ simply translates the spectrum of $x_{BB}(t)$ to the band around $\pm \omega_c$. Thus, the bandwidth of $x_{AM}(t)$ is twice that of $x_{BB}(t)$. The signal $A_c \cos \omega_c t$ is generated by a “local oscillator” (LO).

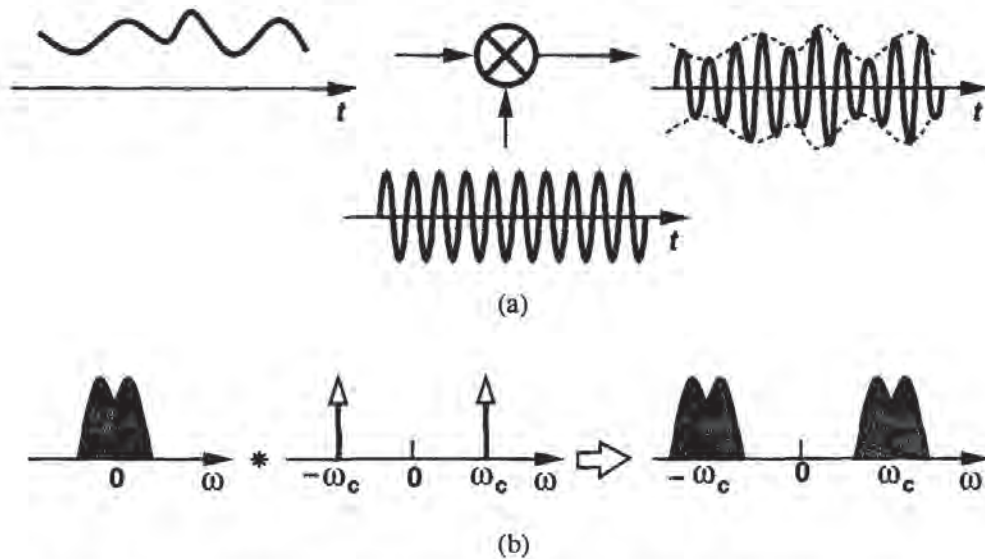


Figure 3.4 Amplitude modulation in (a) time domain, (b) frequency domain.

An AM signal can be demodulated using a multiplier and a low-pass filter. Depicted in Fig. 3.5(a), this technique translates the AM spectrum back to the baseband. Note that here the receiver must ensure the local oscillator generates the same frequency as the carrier frequency of $x_{AM}(t)$.

If the term $1 + m x_{BB}(t)$ remains positive for all t , then the envelope of the waveform shown in Fig. 3.4(a) does not cross zero, and a leaky peak detector can be used to demodulate the signal [Fig. 3.5(b)]. The simplicity of this circuit has made it the dominant choice in AM radios.

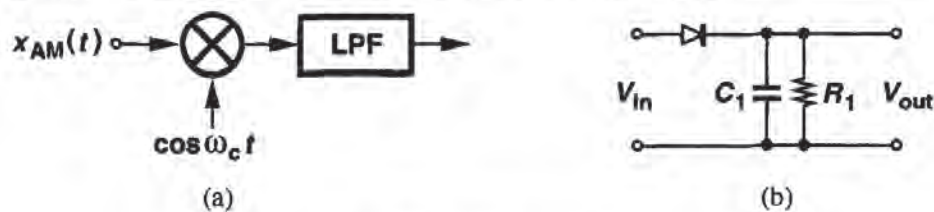


Figure 3.5 AM detectors.

It can be shown that the SNR at the output of both detectors is

$$SNR_{\text{out}} = \frac{A_c^2 m^2 \overline{x_{BB}^2(t)}}{2N_0 B}, \quad (3.2)$$

where $\overline{x_{BB}^2(t)}$ and B denote the mean square value and bandwidth of $x_{BB}(t)$, respectively [1]. In the case of an envelope detector, (3.2) actually holds only for sufficiently high SNR at the input, a reasonable assumption because the quality of reception is acceptable only if the SNR exceeds approximately 25 dB. Note that the transmitted power is equal to $A_c^2[1 + m^2 \overline{x_{BB}^2(t)}]/2$.

Except for broadcast radios and the sound in television, amplitude modulation finds limited use in today's wireless systems. This is because carrying information in the amplitude both makes the signal susceptible to noise and requires a highly linear power amplifier in the transmitter.

3.2.2 Phase and Frequency Modulation

Phase modulation (PM) and frequency modulation (FM) are important concepts that are encountered not only within the context of modems but also in the analysis of such circuits as oscillators and frequency synthesizers.

In Eq. (3.1), if $a(t)$ is a constant, A_c , and the excess phase, $\theta(t)$, is linearly proportional to the baseband signal, we say the carrier is phase modulated, $x_{PM}(t) = A_c \cos[\omega_c t + m x_{BB}(t)]$, where m is the phase modulation index. For example, if $x_{BB}(t) = \alpha t$, then $x_{PM}(t) = A_c \cos(\omega_c + m\alpha)t$, indicating that a ramp baseband waveform simply shifts the carrier frequency by a constant value.

Similarly, if the excess frequency, $d\theta/dt$, is linearly proportional to the baseband signal, we say the carrier is frequency modulated, $x_{FM}(t) = A_c \cos[\omega_c t + m \int_{-\infty}^t x_{BB}(t) dt]$, where m is the frequency modulation index. For example, with a dc baseband signal, $x_{BB}(t) = A$, we have $x_{FM}(t) = A_c \cos(\omega_c + mA)t$. Thus, a dc input to a frequency modulator shifts the carrier frequency by a constant value, a useful result in the context of voltage-controlled oscillators (Chapter 7).

Figure 3.6 shows a baseband signal and the corresponding phase- and frequency-modulated waveforms. Note that in general it is not possible to determine whether a waveform is FM or PM if the baseband signal is unknown. Also, the maximum frequency deviation is equal to $m dx_{BB}(t)/dt$ in PM and $m x_{BB}(t)$ in FM.

The nonlinear dependence of $x_{PM}(t)$ and $x_{FM}(t)$ upon $x_{BB}(t)$ makes it difficult to calculate the bandwidth occupied by the modulated waveforms. We return to this problem later.

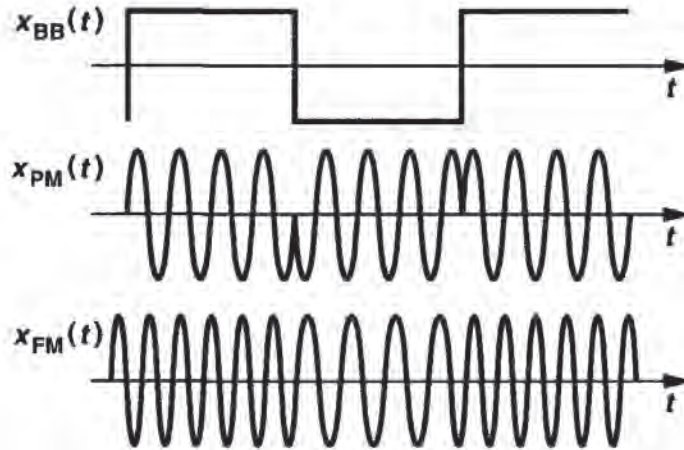


Figure 3.6 Phase- and frequency-modulated waveforms corresponding to a square wave baseband signal.

Analog FM is more widely used than analog PM, perhaps because it is easier to perform frequency modulation and demodulation. Figure 3.7 illustrates a simple frequency modulator, where the resonance frequency of an LC oscillator is varied in proportion to the baseband signal. The voltage-dependent capacitor required here is usually a reverse-biased varactor diode whose junction capacitance varies with $x_{BB}(t)$.

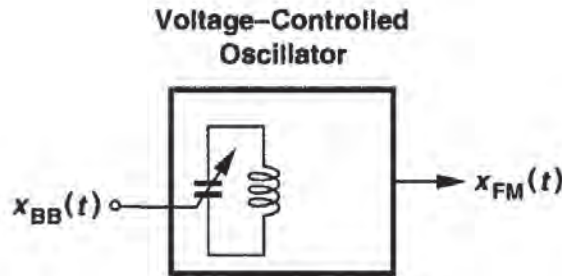


Figure 3.7 Simple frequency modulator.

Frequency demodulation can be performed, in principle, by any circuit whose transfer function is sensitive to frequency. For example, a high-pass filter (i.e., a differentiator) generates an output swing proportional to the signal frequency (Fig. 3.8), in essence converting FM to AM. For $v_{in}(t) = A_c \cos[\omega_c t + m \int x_{BB}(t) dt]$, the output can be approximated as the time derivative of the input:

$$v_{out}(t) \approx A_c R_1 C_1 [\omega_c + m x_{BB}(t)] \sin \left[\omega_c t + m \int x_{BB}(t) dt \right]. \quad (3.3)$$

Thus, if $\omega_c + m x_{BB}(t)$ always remains positive, an envelope detector can be used to recover $x_{BB}(t)$. It is important to note that variations in the *amplitude* of v_{in} corrupt the demodulated waveform, thereby mandating the use of “limiters” before the differentiators.

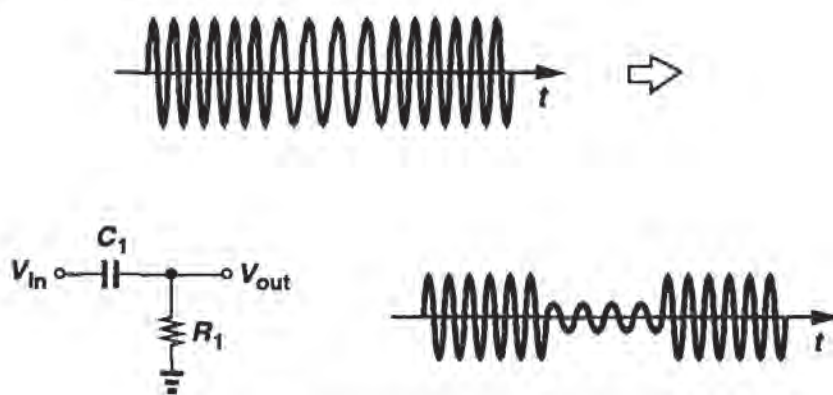


Figure 3.8 Simple frequency demodulator.

The nonlinear dependence of $x_{FM}(t)$ upon $x_{BB}(t)$ makes it difficult to calculate the bandwidth occupied by the modulated waveform. We first consider a simple case called “narrowband FM,” defined as a signal for which $m \int x_{BB}(t)dt \ll 1$ rad. The FM waveform can then be approximated as

$$x_{FM,NB}(t) \approx A_c \cos \omega_c t - A_c m (\sin \omega_c t) \int x_{BB}(t)dt. \quad (3.4)$$

Consequently, as shown in Fig. 3.9(a), the spectrum consists of the carrier and that of $\int x_{BB}(t)dt$ translated to $\pm\omega_c$, indicating that the bandwidth of $x_{FM}(t)$ is somewhat *less* than that of $x_{AM}(t)$ of Fig. 3.4(a) because of the $1/s$ attenuation resulting from integration.

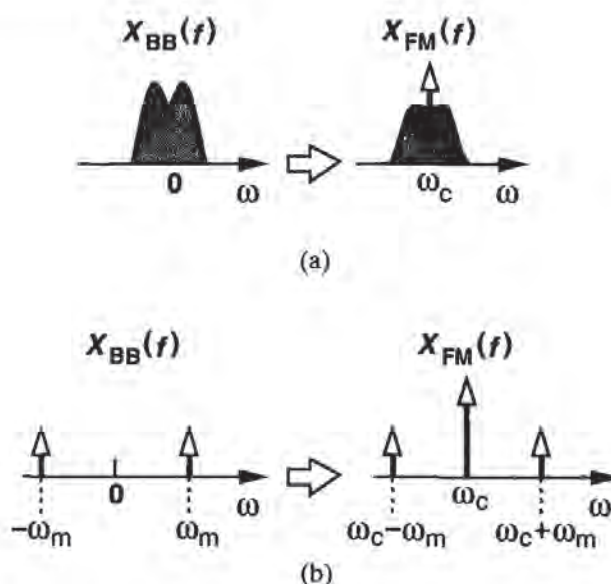


Figure 3.9 Narrowband FM with (a) random and (b) sinusoidal modulation.

A special case of narrowband FM that proves quite useful is for a sinusoidal baseband signal, e.g., $x_{BB}(t) = A_m \cos \omega_m t$. In this case, we have

$$x_{FM,NB}(t) \approx A_c \cos \omega_c t - A_m A_c \frac{m}{\omega_m} \sin \omega_c t \sin \omega_m t \quad (3.5)$$

$$\begin{aligned} &= A_c \cos \omega_c t - \frac{A_m A_c m}{2\omega_m} \cos(\omega_c - \omega_m)t \\ &\quad + \frac{A_m A_c m}{2\omega_m} \cos(\omega_c + \omega_m)t. \end{aligned} \quad (3.6)$$

Illustrated in Fig. 3.9(b), the resulting spectrum consists of impulses at $\pm\omega_c$ and “sidebands” at $\pm(\omega_c \pm \omega_m)$.

Equations (3.5) and (3.6) yield three important points. First, for the narrowband FM assumption to hold, we must have $m A_m / \omega_m \ll 1$ rad. Second, as ω_m increases, the magnitude of the sidebands decreases; i.e., higher modulating frequencies have less weighting in the FM signal. Third, if $x_{FM}(t) = A_c \cos[\omega_c t - (m A_m / \omega_m) \sin \omega_m t]$, then the maximum frequency deviation is equal to $m A_m$. Note that the spacing between the sidebands and the carrier is equal to ω_m and has no relationship with the maximum frequency deviation.

While a useful concept in many communication circuits, narrowband FM has limited application in modems because of its low SNR, an issue described below. As the next step toward a realistic case, we consider an FM signal with a sinusoidal baseband waveform, but without the restriction $m A_m / \omega_m \ll 1$ rad. If $x_{BB}(t) = A_m \cos \omega_m t$, then $x_{FM}(t) = A_c \cos[\omega_c t - (m A_m / \omega_m) \sin \omega_m t]$. This expression can be expanded as

$$x_{FM}(t) = A_c \sum_{n=-\infty}^{+\infty} J_n(\beta) \cos(\omega_c + n\omega_m)t, \quad (3.7)$$

where $J_n(\cdot)$ denotes the n th order Bessel function of the first kind and $\beta = m A_m / \omega_m$. If $\beta \ll 1$ rad (narrowband FM), then $J_0(\beta) \approx 1$, $J_{\pm 1}(\beta) \approx \pm\beta/2$, and $J_n(\beta) \approx 0$ for $|n| > 1$. Thus, $x_{FM}(t) = A_c \cos \omega_c t + (A_c \beta/2) \cos(\omega_c + \omega_m)t - (A_c \beta/2) \cos(\omega_c - \omega_m)t$, the same result as that in (3.6). As β approaches and exceeds unity, the coefficients $J_n(\beta)$ become significant for $|n| > 1$, introducing components at $\omega_c \pm n\omega_m$. Shown in Fig. 3.10 are examples of the FM spectra for $\beta = 1$ and $\beta = 2$. More general cases of FM such as multiple-frequency or Gaussian modulation are treated in [2].

An important conclusion to be drawn from the above analysis is that the bandwidth of FM signals depends on m , A_m , and ω_m . In fact, if the bandwidth, B_{FM} , is defined as that containing 98% of the signal power, then $B_{FM} \approx 2(\beta + 1)B_{BB}$, where B_{BB} is the bandwidth of the baseband signal [1]. This is called Carson’s rule.

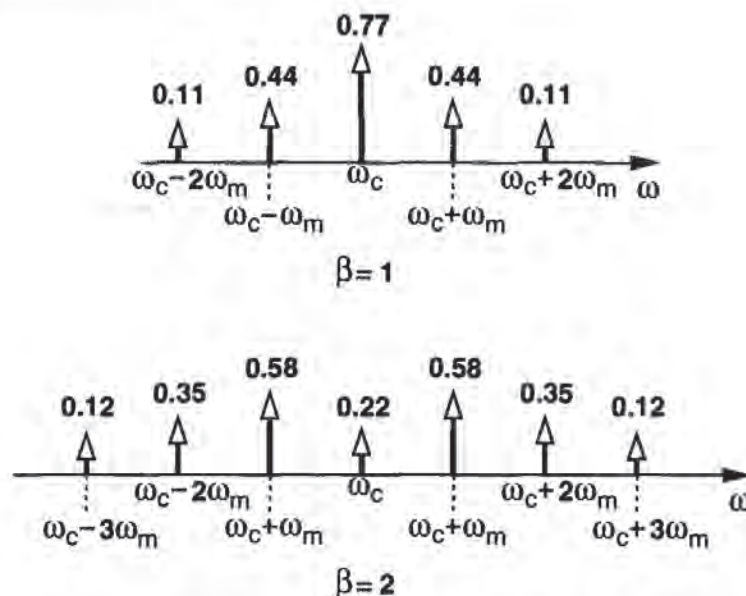


Figure 3.10 FM spectra for two different values of β .

In Eq. (3.6), we noted that as the frequency of the modulating signal increases, the magnitude of the corresponding component in the FM waveform decreases. The FM demodulator thus compensates for this rolloff by introducing larger gain at higher frequencies, as can be seen in the differentiator of Fig. 3.8. However, this leads to amplification of noise at high frequencies, thereby degrading the overall SNR. For this reason, most FM transmitters shape the baseband spectrum by amplifying the high-frequency end ("preemphasis") (Fig. 3.11). In the receiver, the detected signal is applied to a low-pass filter to cancel the preemphasis ("deemphasis"), at the same time suppressing high-frequency noise. This method does not usually increase the required bandwidth significantly if the original baseband spectrum has a low-pass shape.

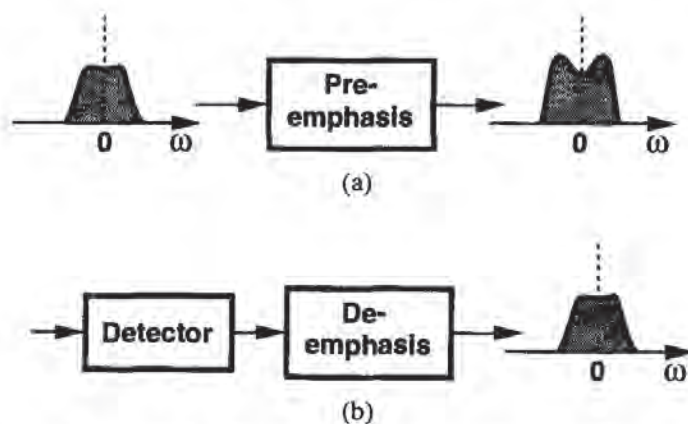


Figure 3.11 Preemphasis and deemphasis in FM.

For an FM detector without preemphasis and deemphasis, it can be shown that

$$\frac{SNR_{out}}{SNR_{in}} = 6\beta^2(\beta + 1) \frac{\overline{x_{BB}^2(t)}}{V_P^2}, \quad (3.8)$$

where V_P is the peak value of $x_{BB}(t)$ [2]. With preemphasis and deemphasis,

$$\frac{SNR_{out}}{SNR_{in}} = 2\beta^2(\beta + 1) \left(\frac{B}{f_1}\right)^2 \frac{\overline{x_{BB}^2(t)}}{V_P^2}, \quad (3.9)$$

where B is the bandwidth of the baseband signal and f_1 is the -3 -dB corner frequency of the low-pass filter used in deemphasis. For typical applications, (3.9) yields a value 10 to 15 dB higher than (3.8) [1]. In both cases, the performance strongly depends on β , revealing a trade-off between signal quality and bandwidth efficiency.

The last attribute of FM systems to be considered is the power efficiency. Since an FM signal has a constant envelope and carries the information in the zero crossings, it can be processed by a nonlinear power amplifier with no corruption of the baseband signal or increase in the required bandwidth. Thus, FM signals are well suited to high-efficiency (inevitably nonlinear) PAs.

3.3 DIGITAL MODULATION

In “digital” RF systems, the carrier is modulated by a digital baseband signal. As explained in Chapter 4, digital modulation offers many advantages over analog modulation and is widely used in wireless systems.

The digital counterparts of AM, PM, and FM, are called amplitude shift keying (ASK), phase shift keying (PSK), and frequency shift keying (FSK), respectively. Figure 3.12 illustrates examples of these waveforms for a binary baseband signal. In RF applications, PSK and FSK find wider usage than does ASK because of their lower sensitivity to amplitude noise.

As mentioned in Section 3.1, analysis of a modem entails quantifying three parameters: signal quality, spectral efficiency, and power efficiency. The quality of digital signaling methods is expressed in terms of the bit error rate (BER), defined as the average number of erroneous bits observed at the output of the detector divided by the total number of bits received in a unit time. Thus, the goal is to calculate the *probability of error* in the presence of noise and other interferers.

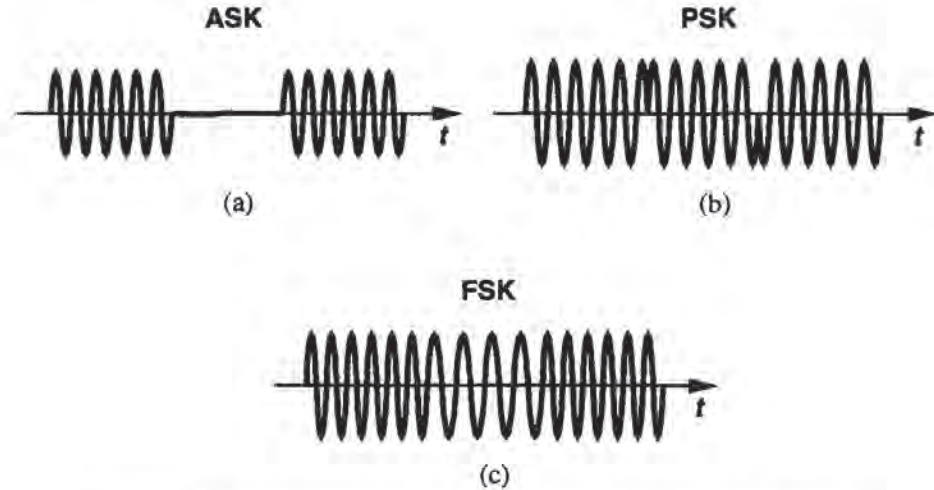


Figure 3.12 (a) Amplitude, (b) phase, and (c) frequency shift keying.

3.3.1 Basic Concepts

Analysis of digital modems requires a number of conceptual tools that are not used in the more familiar analog modulation systems. We briefly review some of the essential concepts here.

Binary and M -ary Signaling A natural choice of the digital baseband signal is a binary waveform, which can be expressed as

$$x_{BB}(t) = \sum_n b_n p(t - nT_b), \quad (3.10)$$

where b_n is the bit value in the time interval $[nT_b, (n+1)T_b]$ and assumes one of two values, e.g., 0 and 1 or -1 and $+1$. For a rectangular pulse shape, the receiver may simply decide whether the (peak) value of the waveform is less or greater than a threshold to detect the bit value.

In practice, it is sometimes advantageous to employ multilevel “digital” signals to modulate the carrier, usually because bandwidth requirements can be relaxed. Shown in Fig. 3.13 is an example where the original binary data stream is subdivided into groups of two bits, and each group is converted to one of four possible levels by means of a D/A converter. We note that the resulting waveform exhibits fewer transitions per unit time but requires higher amplitude resolution in the detector. Called “ M -ary” signaling, such a scheme can also be represented by (3.10) if b_n assumes all of the discrete levels. To avoid confusion with the binary case, each transmitted level here is called a “symbol” rather than a bit. In Fig. 3.13, the “symbol rate” is one-half the original bit rate.

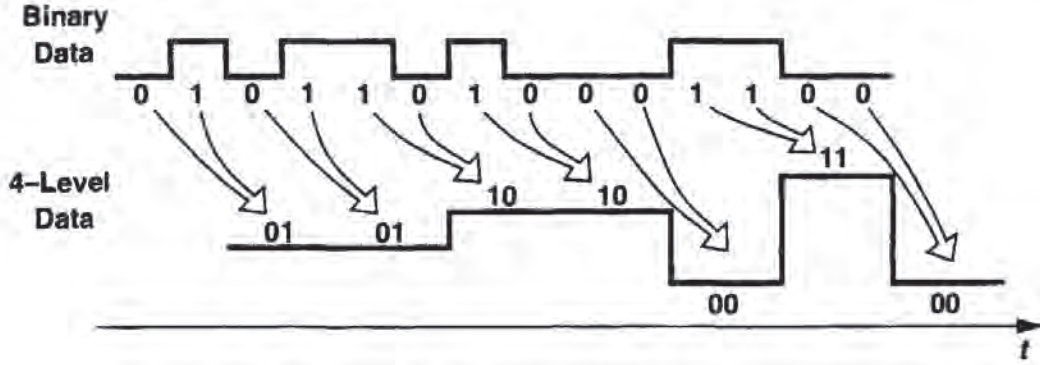


Figure 3.13 Four-level digital representation of a binary data stream.

Basis Functions With the baseband signal described by (3.10), we now consider the representation of digitally modulated waveforms. As an example, the binary FSK signal of Fig. 3.12(c) can be written as

$$x_{FSK}(t) = A_c \cos \omega_1 t, \quad \text{if } b_n = 0, \quad (3.11)$$

$$= A_c \cos \omega_2 t, \quad \text{if } b_n = 1. \quad (3.12)$$

This indicates that every transmitted bit can be represented by an inner product,

$$x_{FSK}(t) = \alpha_1 \phi_1(t) + \alpha_2 \phi_2(t) \quad (3.13)$$

$$= [\alpha_1 \ \alpha_2] \cdot [\phi_1(t) \ \phi_2(t)], \quad (3.14)$$

where $[\alpha_1 \ \alpha_2] = [0 \ A_c]$ or $[A_c \ 0]$, $\phi_1(t) = \cos \omega_1 t$, and $\phi_2(t) = \cos \omega_2 t$. In other words, $x_{FSK}(t)$ is expressed as a linear combination of two functions.

The idea of representing a waveform as a linear combination of “basis functions” should be familiar from Fourier series: a function $x(t)$ (with finite energy) can be written as

$$x(t) = \sum_{n=0}^{\infty} a_n \cos n\omega t + \sum_{n=0}^{\infty} b_n \sin n\omega t, \quad (3.15)$$

over the interval $t_0 < t < t_0 + T$, where $\omega = 1/T$. An important property of the Fourier series is the “orthogonality” of the basis functions, namely,

$$\int_0^T \cos n\omega t \cos m\omega t = 0, \quad (3.16)$$

$$\int_0^T \cos n\omega t \sin m\omega t = 0, \quad (3.17)$$

$$\int_0^T \sin n\omega t \sin m\omega t = 0, \quad (3.18)$$

for $n \neq m$.

With the above observations, we now state that each symbol in a digitally modulated waveform is represented by a linear combination of orthogonal basis functions,²

$$x(t) = \alpha_1 \phi_1(t) + \alpha_2 \phi_2(t) + \cdots + \alpha_N \phi_N(t), \quad (3.19)$$

where

$$\int_0^{T_S} \phi_m(t) \phi_k(t) dt = 0 \quad (3.20)$$

for $m \neq k$ and T_S is the symbol period. The number N is called the “dimension” of the set of basis functions.

The reader may wonder about the motivation behind the foregoing development. We justify this by two reasons. First, expressing waveforms as a linear combination of functions allows a unified treatment and fair comparison of many seemingly different modulation schemes. Second, orthogonal basis functions form a “minimum set” for representing waveforms, and, more importantly, they can be used in optimum detection of signals.

Signal Constellations For dimensions below 4, we can visualize the modulated waveforms in terms of the coefficients in their inner products. That is, in $x(t) = [\alpha_1 \ \alpha_2 \ \cdots] \cdot [\phi_1(t) \ \phi_2(t) \ \cdots]$, we simply ignore the basis functions and plot all possible values of the vector $[\alpha_1 \ \alpha_2 \ \cdots]$ in cartesian coordinates. Thus, for the binary FSK signal of Fig. 3.12(c), we obtain the plot shown in Fig. 3.14(a). Called a “signal constellation” or a “signal space,” such a plot proves valuable in understanding various effects in digital modems.

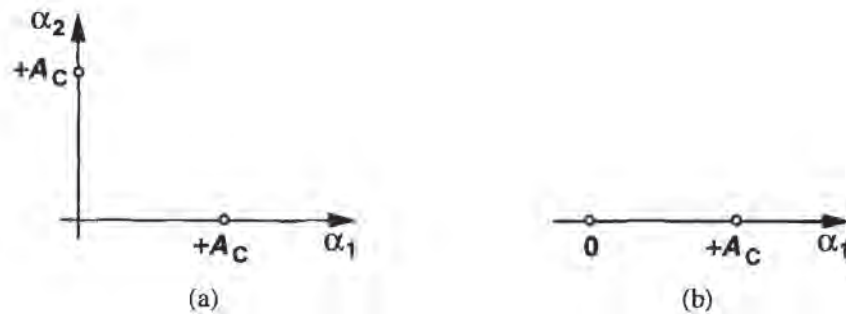


Figure 3.14 Signal constellations of (a) FSK, (b) ASK.

² The basis functions are usually normalized so as to have unit energy over T_b , but here we ignore this detail.

As another example, consider the binary ASK waveform of Fig. 3.12(a), where the modulated signal corresponding to each bit is expressed as

$$x_{ASK}(t) = A_c \cos \omega_c t, \quad \text{if } b_n = 1, \quad (3.21)$$

$$= 0, \quad \text{if } b_n = 0. \quad (3.22)$$

Consequently, $x_{ASK}(t) = \alpha_1 \phi_1(t)$, where α_1 assumes a value of 0 or A_c and $\phi_1(t) = \cos \omega_c t$. This scheme has one dimension, with the constellation shown in Fig. 3.14(b).

An immediate application of signal constellations is in understanding the effect of noise upon detection of the received waveform. Suppose as shown in Fig. 3.15, a binary ASK signal is corrupted by some noise, $n(t)$, and the result is applied to an envelope detector. The signal thus obtained resembles the original baseband waveform, but with substantial amplitude noise. Therefore, an analog comparator with a threshold at midway between the expected high and low levels is added to sense the transitions and generate a less noisy output.³



Figure 3.15 Effect of noise on (a) an ASK signal and (b) its constellation.

What happens if the rms value of the noise is increased? If the noise amplitude is large enough, the output of the envelope detector occasionally crosses the threshold erroneously, producing an incorrect bit value at the output of the comparator. In other words, as the noise amplitude becomes comparable with half the difference between the high and low levels, the probability of error rises rapidly.

To see the effect of noise on the constellation, we examine the levels at the output of the envelope detector for a long time, obtaining a “noise band” around each point [Fig. 3.15(b)]. A greater noise amplitude widens the variations of the detected level, thereby increasing the probability of error, whereas a greater difference between the two points (e.g., larger signal amplitude) decreases this probability. Thus, the cartesian distance between the points in a constellation is an indication of the robustness of the modulated signal to noise. In fact, the points with minimum distance have the most significant impact on the bit error rate. We also see intuitively that if the two levels in Fig. 3.15(b) occur with equal probability, then the decision threshold should be at $A_c/2$.

³ This detection technique is not optimum, but used here to illustrate a point.

As another example, consider the binary (orthogonal) FSK constellation of Fig. 3.14(a) in the presence of noise. The resulting signal space appears as shown in Fig. 3.16. As the noise power increases, the “clouds” extend beyond the decision boundary, thereby causing errors in the detector.

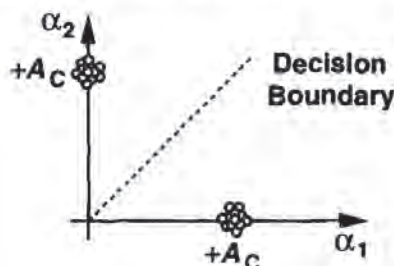


Figure 3.16 Effect of noise on FSK constellation.

Optimum Detection The manner in which the baseband signal is extracted from the modulated waveform has great impact on the overall performance of the system, in particular the signal quality. Since the baseband signal is digital, the detector output must be sampled every bit period, T_b , to determine the received value. This is illustrated in Fig. 3.17.

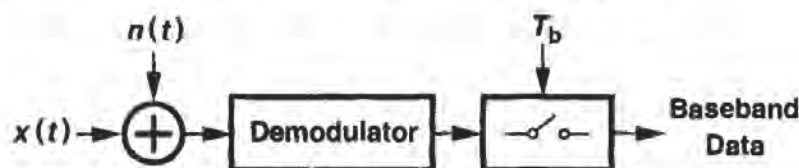


Figure 3.17 Signal detection by sampling.

To understand the problem of optimum detection, we begin with the simple case depicted in Fig. 3.18. Suppose a single pulse (not necessarily modulated) is to be detected in the presence of additive noise. How should the pulse be shaped so as to maximize the signal-to-noise ratio at the sampling instant? We may surmise that if $h(t) = \delta(t)$ and the sampling is synchronized such that the peak value of the pulse is sensed, the output SNR is high [Fig. 3.19(a)].

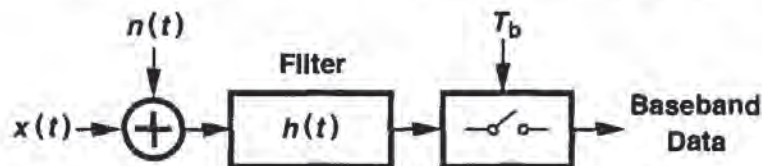


Figure 3.18 Use of filter in detector.

However, we can take advantage of the random nature of noise by performing averaging in one bit period. Shown in Fig. 3.19(b) is an example where the input pulse is integrated from 0 to T_b and the sampling occurs at $t = T_b$. In this case, noise components that vary significantly in a period of T_b tend to average out.

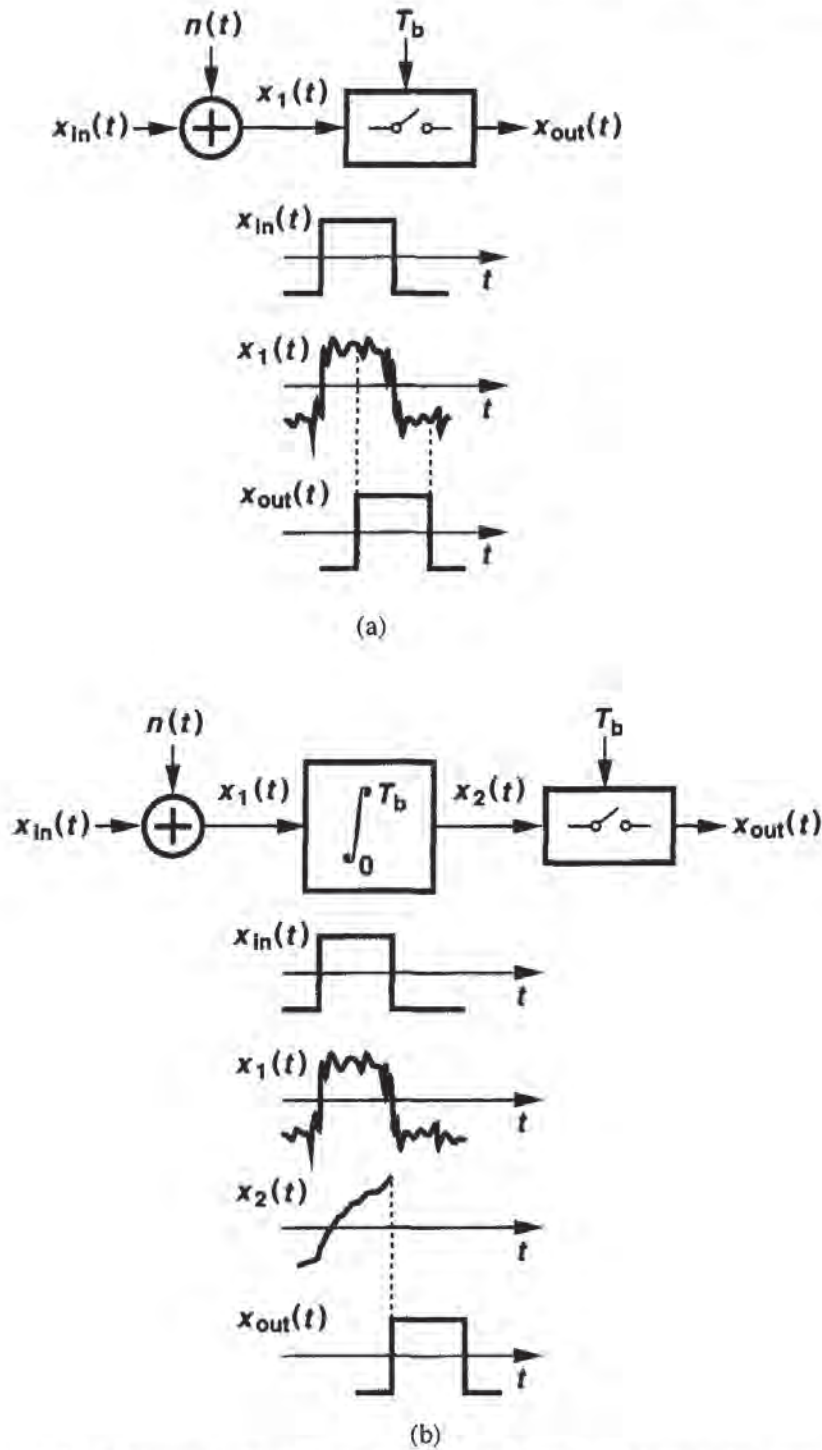


Figure 3.19 Detector with (a) sampling the peak value, (b) integration over one bit prior to sampling.

The example of Fig. 3.19(b) leads us to the concept of “matched filter.” It can be shown that for a pulse $p(t)$ that is corrupted by *additive white noise*, there exists an optimum filter that maximizes the SNR at the sampling instant [1].

Called a matched filter (Fig. 3.20), such a circuit has an impulse response given by $h(t) = p^*(T_b - t)$, where the right-hand side is the conjugate of $p(t)$ reversed in time and shifted to the right by T_b . In most cases of interest, $p(t)$ is real and $h(t) = p(T_b - t)$. Note that the magnitude of the Fourier transform of $h(t)$ is equal to that of $p(t)$.

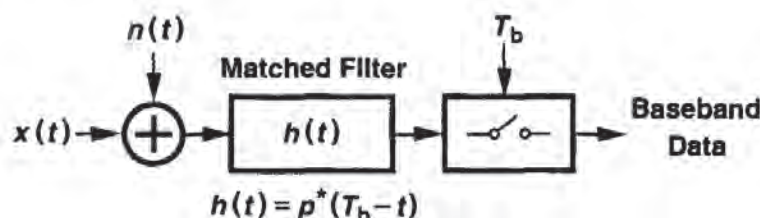


Figure 3.20 Detection using matched filtering.

To gain an intuitive feeling about matched filters, let us assume $p(t)$ is a rectangular pulse as shown in Fig. 3.21. In the absence of noise, the filter output is equal to

$$y(t) = p(t) * h(t) \quad (3.23)$$

$$= \int_{-\infty}^{+\infty} p(t - \tau)h(\tau)d\tau, \quad (3.24)$$

that is, a triangular pulse with a maximum value at $t = T_b$. We note that the impulse response is that of an integrator (with the output “dumped” at $t = T_b$.) This confirms our earlier conviction about averaging the noise.

If the input to a matched filter consists of a pulse $p(t)$ and additive white noise with power density $N_0/2$, then the maximum SNR is equal to [1]

$$SNR_{\max} = \frac{2E_p}{N_0}, \quad (3.25)$$

where

$$E_p = \int_{-\infty}^{+\infty} |p(t)|^2 dt. \quad (3.26)$$

We say E_p is the energy of the signal. If $p(t)$ is a voltage quantity, E_p is the energy dissipated in a 1- Ω resistor driven by $p(t)$. Equation (3.25) yields two interesting results. First, $SNR_{\max}$ depends on the energy of the pulse but not its shape. Second, $SNR_{\max}$ does not depend on the bandwidth occupied by the pulse or the matched filter.

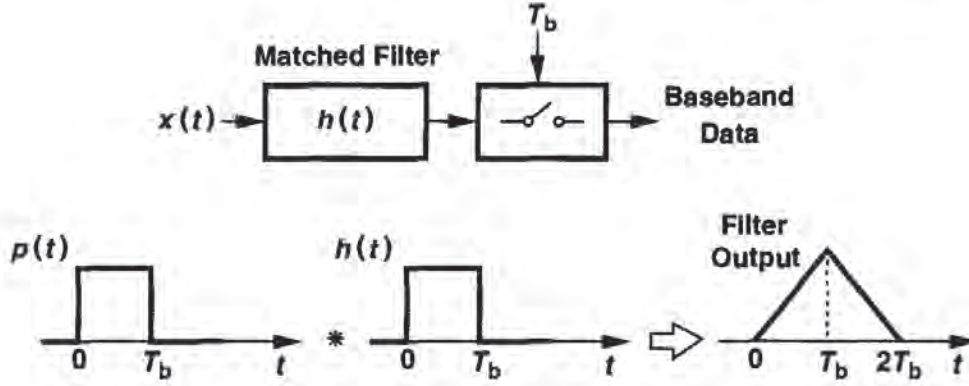


Figure 3.21 Optimum detection of a rectangular pulse.

With the above development, we are now ready to study the optimum detection of *modulated* signals. We first recognize that if the input waveform corresponding to a bit is a real function, $p(t)$, then the matched filter output at the end of the bit period can be written as

$$y(T_b) = \int_{-\infty}^{+\infty} x(\tau) h(T_b - \tau) d\tau \quad (3.27)$$

$$= \int_{-\infty}^{+\infty} x(\tau) p(\tau) d\tau, \quad (3.28)$$

where $x(t) = p(t) + n(t)$. If $p(t)$ is zero outside the interval $[0, T_b]$, then

$$y(T_b) = \int_{\tau=0}^{\tau=T_b} x(\tau) p(\tau) d\tau. \quad (3.29)$$

Equation (3.29) is an important result, suggesting that a “correlation function” must be performed between the input (including both the signal and the noise) and the known pulse shape, $p(t)$, for one bit period. Illustrated in Fig. 3.22, this technique reduces to the topology of Fig. 3.21 if $p(t)$ is a rectangular pulse.

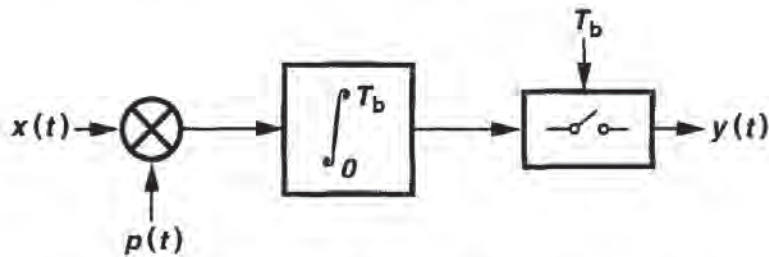


Figure 3.22 Optimum detection using a correlator.

The function $p(t)$ can assume any shape, including the modulated waveforms of Fig. 3.12. But it is important to synchronize the starting point of the pulse in $x(t)$ with that of $p(t)$, because timing misalignment results in a smaller peak value at the correlator output. This issue is resolved by “bit (or symbol) synchronization.”

Now recall that a digitally modulated waveform can be expressed as a linear combination of the basis functions. To detect the value of the transmitted level, we can generalize the idea of the correlator. Shown in Fig. 3.23 is an example for a two-dimensional signal space, where the correlation between the input and each basis function is computed and the results are properly combined to generate the baseband waveform. The operation of this receiver is more intuitively understood if we consider the familiar case of binary FSK, shown in Fig. 3.24. Here, each correlator serves to find how “similar” $x_{FSK}(t)$ is to $\cos \omega_1 t$ or $\cos \omega_2 t$ by integrating $x_{FSK}(t) \cos \omega_1 t$ and $x_{FSK}(t) \cos \omega_2 t$ over one bit period. The adder then decides which sampled value is greater, generating the output binary value accordingly.

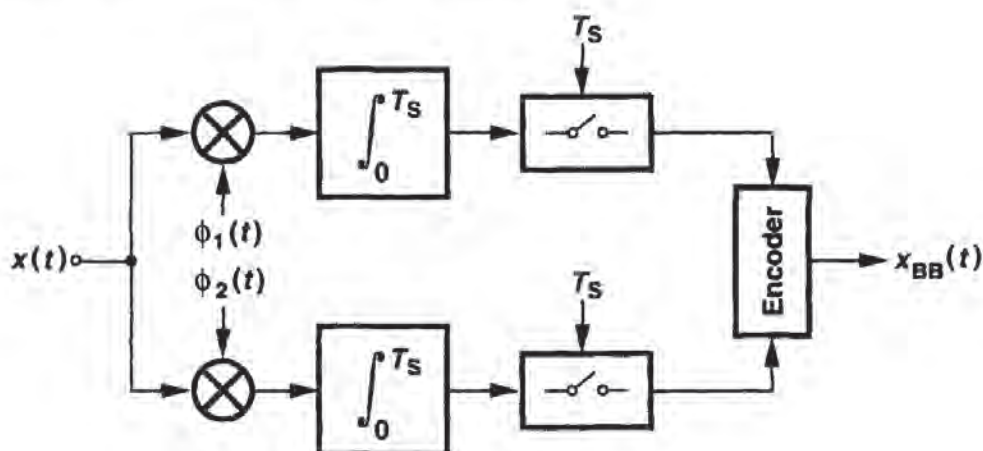


Figure 3.23 Correlation receiver for two-dimensional signal space.

Coherent and Noncoherent Detection The binary FSK receiver of Fig. 3.24 can be used to distinguish between two methods of detection. First, note that the basis functions $\cos \omega_1 t$ and $\cos \omega_2 t$ are generated in the receiver, for example, by means of two oscillators, whereas the waveform $x_{FSK}(t)$ is synthesized in the transmitter. Now suppose, due to start-up phase uncertainty, one of the oscillators actually produces $\sin \omega_1 t$ while the received waveform is $A_c \cos \omega_1 t$. The correlation of these signals over one bit period is equal to zero if T_b is an integer multiple of $2\pi/\omega_1$. The receiver thus fails to generate a correct output. This example indicates the need for *phase synchronization* between the carrier of the received signal and the oscillator output in the receiver. Detection schemes that require phase synchronization are called “coherent.”

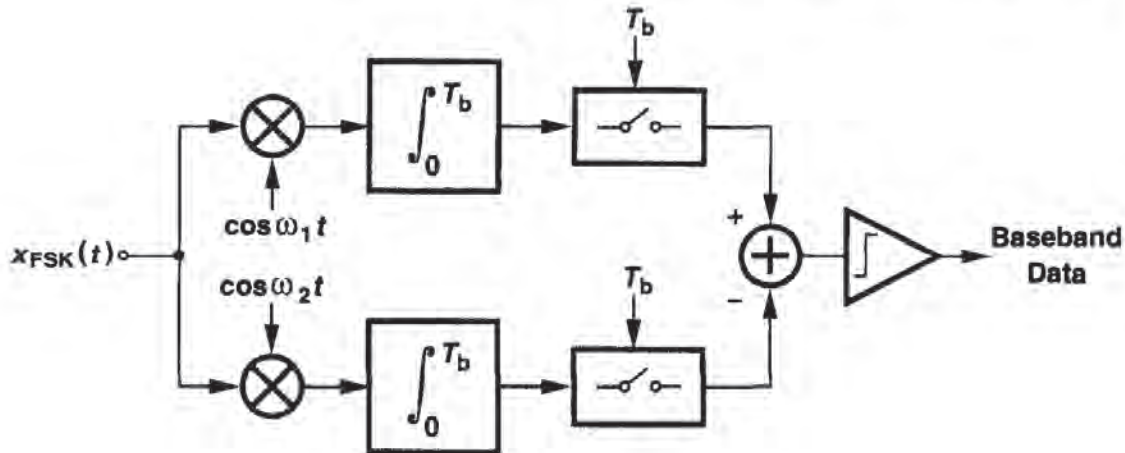


Figure 3.24 Coherent FSK detector.

By contrast, some modulated waveforms can be demodulated using “non-coherent” detection. A simple example is the FM detector of Fig. 3.8. Another example is the binary FSK detector illustrated in Fig. 3.25. This circuit employs two narrowband filters centered at ω_1 and ω_2 followed by envelope detectors to determine which frequency is received during a bit period.

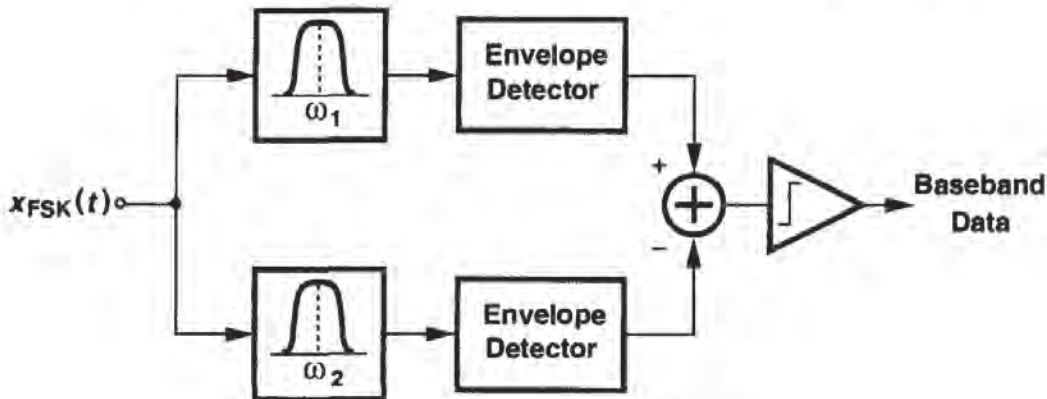


Figure 3.25 Noncoherent FSK detector.

Coherent detectors are usually based on the matched filter concept, providing a lower bit error rate than do their noncoherent counterparts. However, the latter are more widely used in RF design owing to their lesser complexity (Section 3.5)

Definition of Bandwidth In our study of analog FM, we noted that 98% of the signal power lies in a bandwidth equal to $2(\beta + 1)B$. In digital modulation, it is common practice to specify the “99% bandwidth,” namely, that containing 99% of the signal power. Although somewhat arbitrary, the choice of this percentage must be consistent for all modulation schemes. Alternatively, the signal bandwidth can be defined in terms of the relative amount of power that appears in an *adjacent* channel. Illustrated in Fig. 3.26, the idea is that the

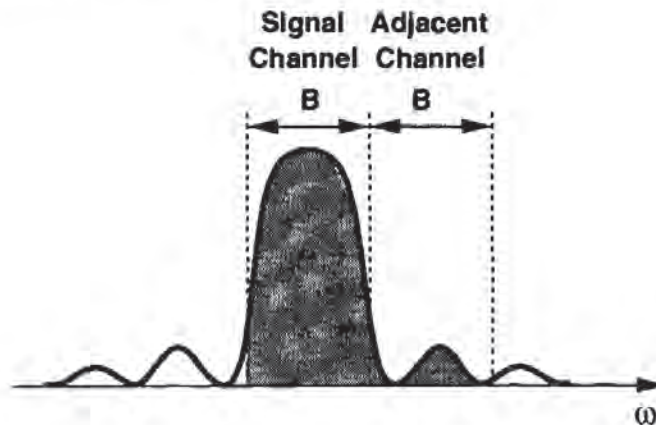


Figure 3.26 Adjacent channel power.

definition of the bandwidth serves to show how closely the signal channels can be spaced, i.e., how much each channel is corrupted by adjacent channels. For example, we say for a bandwidth of 30 kHz, the signal exhibits an “adjacent channel power” (ACP) of -50 dBc, meaning that the power in the adjacent 30-kHz channel divided by the power in the channel carrying the modulated signal is equal to -50 dB.

3.3.2 Binary Modulation

Digital modulation with binary baseband waveforms can be performed as ASK, PSK, or FSK. Since ASK is rarely used in RF applications, we consider only binary PSK (BPSK) and binary FSK (BFSK) here. The goal is to study the bit error rate, spectral efficiency, and power efficiency of each scheme. We assume the received signal is corrupted by additive white Gaussian noise and find the BER of coherent and, if possible, noncoherent detectors in each case.

Before considering any specific modulation, we analyze a general correlator for waveforms modulated by equal-probability binary levels. In contrast to the simple correlator of Fig. 3.22, here we assume *two* possible (real) pulse shapes $p_1(t)$ and $p_2(t)$ to represent binary data. This is because in some cases a logical ZERO is not transmitted as a simple “low” level. For example, in BFSK, $p_1(t) = A_c \cos \omega_1 t$ and $p_2(t) = A_c \cos \omega_2 t$ for $0 < t < T_b$. Shown in Fig. 3.27(a), the optimum receiver thus correlates the input with each of these pulses and compares the results to determine the bit.⁴ This configuration can be simplified as in Fig. 3.27(b), revealing that the equivalent filter is matched to $p_1(t) - p_2(t)$. Therefore, from (3.25)

$$SNR_{\max} = \frac{2E_d}{N_0}, \quad (3.30)$$

⁴ Note that $p_1(t)$ and $p_2(t)$ are not necessarily orthogonal in this case.

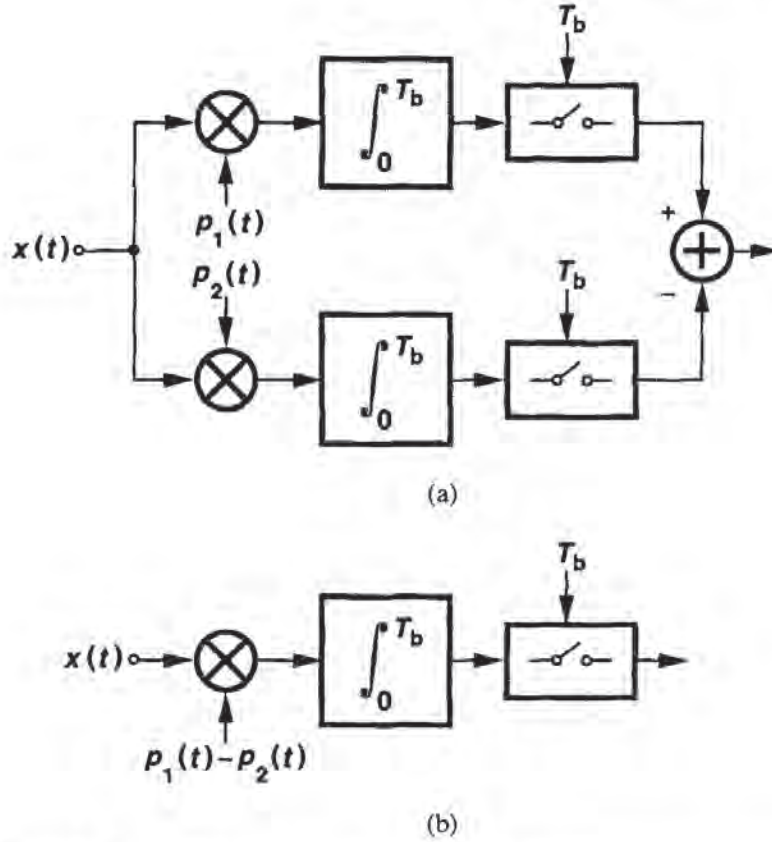


Figure 3.27 (a) Coherent binary receiver, (b) simplified version of (a).

where

$$E_d = \int_{-\infty}^{+\infty} [p_1(t) - p_2(t)]^2 dt. \quad (3.31)$$

Also, it can be shown that E_d is maximum if $p_1(t) = -p_2(t)$.

In the absence of noise, the output of the integrator in Fig. 3.27(b) at $t = T_b$ assumes one of two distinct values, A_1 and A_2 (e.g., $A_1 = +1$ and $A_2 = -1$), corresponding to $x(t) = p_1(t)$ and $x(t) = p_2(t)$, respectively. The decision threshold is set at $(A_1 + A_2)/2$. In the presence of noise, on the other hand, the sampled value is equal to $A_1 + n(T_b)$ or $A_2 + n(T_b)$, where $n(T_b)$ is the noise amplitude at the output of the integrator at $t = T_b$. Note that $n(T_b)$ may not be white, but it is still Gaussian because of the linearity of the correlator.

Now recall from probability theory that if two independent random variables are added, their probability density functions are convolved. Consequently, as illustrated in Fig. 3.28, the quantities $A_1 + n(T_b)$ and $A_2 + n(T_b)$ exhibit two overlapping Gaussian distributions, indicating that an error occurs if $A_1 + n(T_b)$ exceeds the threshold and $p_1(t)$ is transmitted or if $A_2 + n(T_b)$ falls below the threshold and $p_2(t)$ is transmitted. Since the probability that

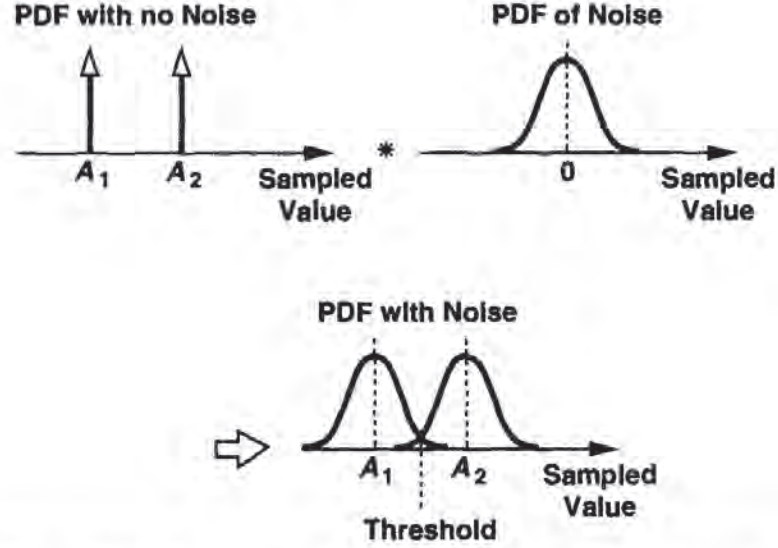


Figure 3.28 Calculation of PDF for binary data with additive noise.

$p_1(t)$ is transmitted is equal to $1/2$, the probability of the first error event is

$$P_{e1} = \frac{1}{2} \int_{(A_1+A_2)/2}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_n} \exp \frac{-(u - A_1)^2}{2\sigma_n^2} du, \quad (3.32)$$

where σ_n is the standard deviation of $n(t)$. The second error event has the same probability, and hence the total probability of error is $P_e = 2P_{e1}$.

With a change of variable, $v = (u - A_1)/\sigma_n$, we have

$$P_e = \int_{(A_2-A_1)/(2\sigma_n)}^{\infty} \frac{1}{\sqrt{2\pi}} \exp \frac{-v^2}{2} dv \quad (3.33)$$

$$= Q\left(\frac{A_2 - A_1}{2\sigma_n}\right), \quad (3.34)$$

where $Q(\cdot)$ is

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} \exp \frac{-u^2}{2} du. \quad (3.35)$$

While not available in closed form, this function can be approximated as

$$Q(x) \approx \frac{1}{x\sqrt{2\pi}} \exp \frac{-x^2}{2}, \quad (3.36)$$

for $x > 3$. Note that the result in (3.34) holds for *any* filter, although $(A_2 - A_1)/(2\sigma_n)$ depends on the type of the filter.

The argument of the Q function in (3.34) is the ratio of two quantities (signal and noise) at the *output* of the integrator. It is more convenient to express this ratio in terms of signal and noise at the *input* of the detector. This is possible for a correlator (matched filter) implementation if we make two observations. First, linear superposition implies that the quantity $A_2 - A_1$

the response of the receiver to an input $x(t) = p_1(t) - p_2(t)$ (although such an input does not occur in actual transmission). Second, since the filter (or correlator) is matched to $p_1(t) - p_2(t)$, if $x(t) = p_1(t) - p_2(t)$ is applied, the SNR at the sampling instant is equal to $(A_2 - A_1)^2/\sigma_n^2$, which reaches a maximum given by (3.30). Thus, $(A_2 - A_1)^2/\sigma_n^2 = 2E_d/N_0$ and, hence,

$$P_e = Q\left(\sqrt{\frac{E_d}{2N_0}}\right). \quad (3.37)$$

Equation (3.37) is an important result, suggesting that the error rate depends on only the (difference) signal energy and the noise spectral density. We emphasize the conditions under which (3.37) holds: additive white Gaussian noise channel and matched-filter (coherent) detection.

The derivations considered thus far apply to all binary modulation schemes. We now study the implications of the above analyses in specific modems.

BPSK In BPSK, the binary baseband data selects one of two opposite phases of the carrier. Illustrated in Fig. 3.29(a), the modulated signal can be written as $x_{BPSK}(t) = A_c \cos(\omega_c t + \phi)$, where $\phi = 0$ or 180° . To obtain the constellation, we write $x_{BPSK}(t) = \alpha \cos \omega_c t$, where $\alpha = +A_c$ or $-A_c$ [Fig. 3.29(b)]. Since $p_1(t) = -p_2(t)$, the correlating signal in the detector is $p_1(t) - p_2(t) = 2p_1(t) = 2A_c \cos \omega_c t$ [Fig. 3.29(c)]. Thus, in the absence of noise, the output of the integrator at $t = T_b$ is equal to

$$V_{\text{int}} = \int_0^{T_b} \pm 2A_c^2 \cos^2 \omega_c t dt. \quad (3.38)$$

For large ω_c , $V_{\text{int}} \approx \pm A_c^2 T_b$, yielding the baseband binary data.

To calculate the error rate with the aid of the results in (3.31) and (3.37), we recognize that

$$E_d = \int_{-\infty}^{+\infty} [p_1(t) - p_2(t)]^2 dt \quad (3.39)$$

$$= \int_0^{T_b} (2A_c \cos \omega_c t)^2 dt \quad (3.40)$$

$$= 2A_c^2 T_b, \quad (3.41)$$

and hence

$$P_e = Q\left(\sqrt{\frac{A_c^2 T_b}{N_0}}\right). \quad (3.42)$$

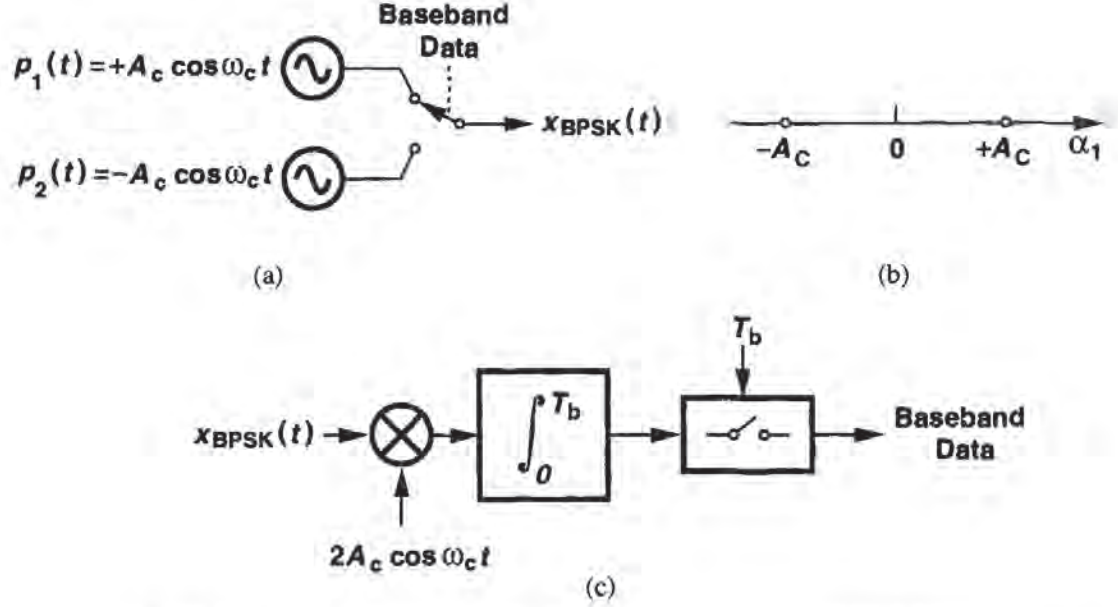


Figure 3.29 BPSK (a) modulator, (b) constellation, and (c) demodulator.

To make fair comparisons with other modulation schemes, we define the (average) energy per bit as $E_b = A_c^2 T_b / 2$, thus obtaining

$$P_{e,BPSK} = Q\left(\sqrt{\frac{2E_b}{N_0}}\right). \quad (3.43)$$

Note that E_b contains both the amplitude and the period of the signal, revealing that the error rate can be lowered by increasing the signal power or decreasing the data rate. The quantity E_b/N_0 is an indication of the signal-to-noise ratio in the receiver.

We find the spectrum of a BPSK waveform by making two observations. First, the modulator of Fig. 3.29(a) can be modified as shown in Fig. 3.30, where the baseband data toggles randomly between $+1$ and -1 and is multiplied by $A_c \cos \omega_c t$:

$$x_{BPSK} = x_{BB}(t) \cdot A_c \cos \omega_c t. \quad (3.44)$$

Thus, the spectrum of $x_{BPSK}(t)$ is simply that of $x_{BB}(t)$ translated to $\pm \omega_c$. Second, a random binary waveform, $x(t)$, employing pulse shapes $+p(t)$ and $-p(t)$ to represent ONES and ZEROS, respectively, has the following power spectral density,

$$S_x(\omega) = \frac{1}{T_b} |P(\omega)|^2, \quad (3.45)$$

where $P(\omega)$ is the Fourier transform of $p(t)$ [1]. If $p(t)$ is a rectangular pulse with duration T_b and unity amplitude, then $P(\omega) = [2 \sin(\omega T_b / 2)] / \omega$, and

$$S_{BPSK}(\omega) = \frac{A_c^2}{T_b} \frac{\sin^2[(\omega + \omega_c)T_b / 2]}{(\omega + \omega_c)^2} + \frac{A_c^2}{T_b} \frac{\sin^2[(\omega - \omega_c)T_b / 2]}{(\omega - \omega_c)^2}. \quad (3.46)$$

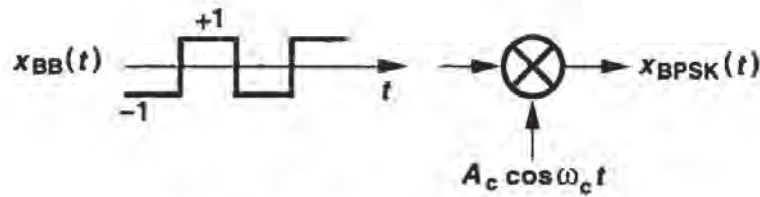


Figure 3.30 Simplified BPSK modulator.

The power efficiency of BPSK is discussed in Section 3.4.

BFSK In BFSK, the binary baseband data selects one of two carrier frequencies with equal amplitudes [Fig. 3.31(a)].⁵ The modulated signal is written as $x_{BFSK}(t) = \alpha_1 \cos \omega_1 t + \alpha_2 \cos \omega_2 t$, where $[\alpha_1 \alpha_2] = [0 A_c]$ or $[A_c 0]$. For the two basis functions to be orthogonal over one bit period, we must have

$$\int_0^{T_b} \cos \omega_1 t \cos \omega_2 t dt = 0. \quad (3.47)$$

For $\omega_1 + \omega_2 \gg \omega_1 - \omega_2$, the above equation reduces to $[\sin(\omega_1 - \omega_2)T_b]/(\omega_1 - \omega_2) = 0$. Thus, $(\omega_1 - \omega_2)T_b = n\pi$, and for minimum spacing between ω_1 and ω_2 , $\omega_1 - \omega_2 = \pi/T_b$ or $f_1 - f_2 = 1/(2T_b)$.

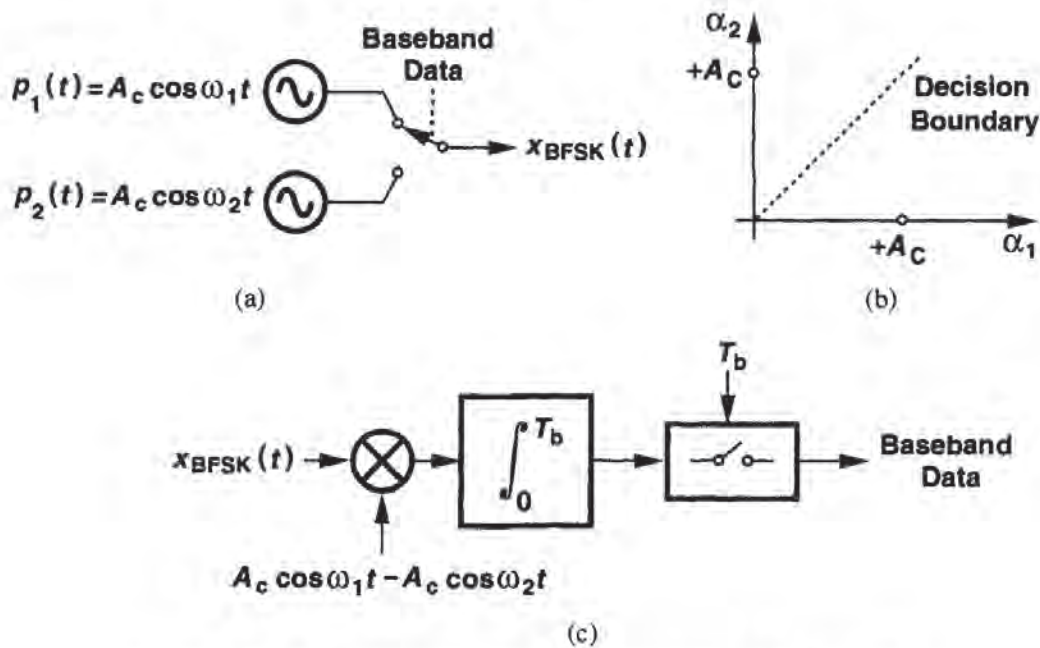


Figure 3.31 BFSK (a) modulation, (b) constellation, (c) detection.

⁵ This approach is called "discontinuous" FSK because the phase changes abruptly.

In contrast to BPSK, orthogonal BFSK is a two-dimensional scheme, exhibiting the constellation in Fig. 3.31(b). Here, $p_1(t) = A_c \cos \omega_1 t$ and $p_2(t) = A_c \cos \omega_2 t$; that is, $p_1(t) \neq -p_2(t)$. As we will see below, this yields interesting differences between the error behavior of BPSK and BFSK.

Shown in Fig. 3.31(c), a BFSK detector correlates the input with $p_1(t) - p_2(t) = A_c \cos \omega_1 t - A_c \cos \omega_2 t$. Consequently, the output of the integrator at $t = T_b$ in the absence of noise is equal to $A_c^2 T_b / 2$ for $x_{BFSK} = A_c \cos \omega_1 t$ and $-A_c^2 T_b / 2$ for $x_{BFSK} = A_c \cos \omega_2 t$.

The bit error rate for coherent BFSK detection is calculated with the aid of (3.31) and (3.37). Since $p_1(t)$ and $p_2(t)$ are orthogonal,

$$E_d = \int_0^{T_b} [p_1^2(t) + p_2^2(t)] dt \quad (3.48)$$

$$= A_c^2 T_b. \quad (3.49)$$

Thus,

$$P_e = Q \left(\sqrt{\frac{A_c^2 T_b}{2N_0}} \right). \quad (3.50)$$

As with BPSK, we define the average energy per bit as $E_b = A_c^2 T_b / 2$, thereby obtaining

$$P_{e,BFSK} = Q \left(\sqrt{\frac{E_b}{N_0}} \right). \quad (3.51)$$

Comparison of (3.43) and (3.51) indicates that, for a given probability of error and noise density, the bit energy in BFSK must be twice that in BPSK. What is the reason for this difference? From the constellations of Figs. 3.29(b) and 3.31(b), we observe that for a given value of carrier amplitude, A_c , the amount of noise that can corrupt a bit so much that the correlator output crosses the threshold erroneously is *higher* in BPSK than in BFSK. In other words, the minimum distance between the points in the constellation is greater in BPSK. From another point of view, recall that in $SNR_{\max} = 2E_d/N_0$, the value of E_d reaches its maximum if $p_1(t) = -p_2(t)$, which is the case for BPSK but not for BFSK.

The above difference is often stated as “BPSK has a 3-dB advantage over BFSK.” Nevertheless, BFSK is widely used in low data rate applications (such as pagers) where E_b can be maximized by allowing a long T_b . The popularity of BFSK stems from the simplicity of its detection, especially in noncoherent systems (Section 3.5), and its power efficiency (Section 3.4).

The spectrum of FSK signals is generally difficult to calculate [1]. From Carson’s rule, the 98% bandwidth is approximately equal to $B_T = 2(\Delta f + 1/T_b)$, where $\Delta f = f_1 - f_2$ [1]. In orthogonal FSK, $\Delta f = 1/(2T_b)$ and $B_T \approx 3/T_b$.

The power efficiency of BFSK is discussed in Section 3.4.

3.3.3 Quadrature Modulation

In many applications, it is beneficial to subdivide a binary data stream into pairs of two bits and represent each pair with one of four levels before performing modulation. For example, bits b_m and b_{m+1} can be impressed upon a *single* carrier as

$$x(t) = b_m A_c \cos \omega_c t - b_{m+1} A_c \sin \omega_c t. \quad (3.52)$$

This is possible because $\cos \omega_c t$ and $\sin \omega_c t$ are orthogonal functions.

Called “quadrature modulation” or “quadrature multiplexing,” this operation is illustrated in Fig. 3.32, where a serial-to-parallel (S/P) converter separates consecutive bits into one stream for the upper arm and another for the lower arm. Since each group of two bits (one in each arm) constitutes a symbol, the symbol rate is half the bit rate (Section 3.3.1), indicating that the required bandwidth is half that of BPSK. This is the principal reason for the wide use of quadrature modulation.

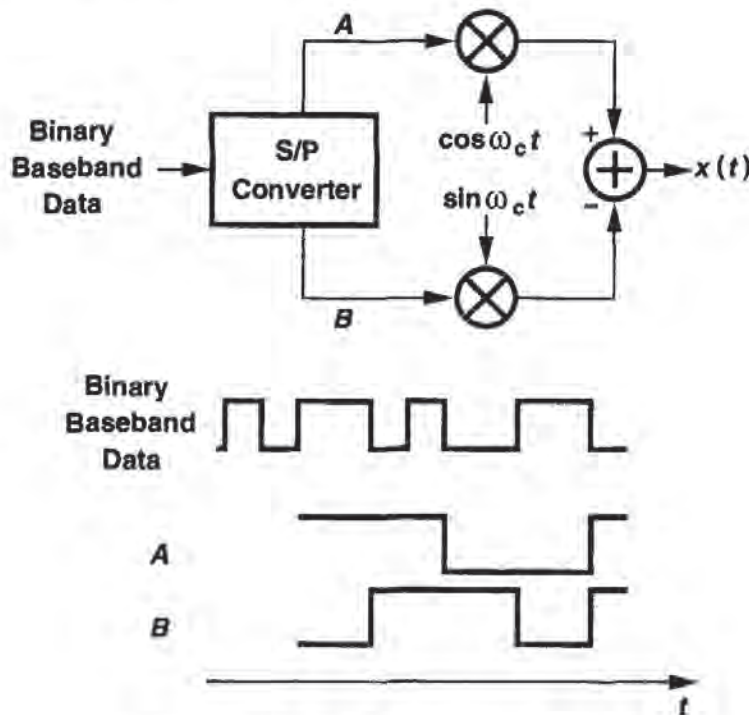


Figure 3.32 Quadrature modulation.

To obtain the constellation, we assume bits b_m and b_{m+1} are rectangular pulses with a height ± 1 and write the modulated signal as $x(t) = \alpha_1 \cos \omega_c t + \alpha_2 \sin \omega_c t$, where α_1 and α_2 can each take on a value of $+A_c$ or $-A_c$. The constellation is shown in Fig. 3.33.

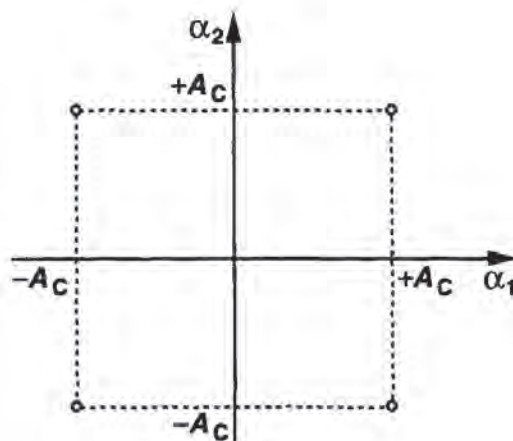


Figure 3.33 Signal constellation for quadrature modulation.

Quadrature modulation encompasses two broad categories: quadrature phase shift keying ($\pi/4$ -QPSK) and minimum shift keying (MSK). The first includes more specific types such as offset QPSK (OQPSK) and $\pi/4$ -QSPK, and the second has a widely used subset, Gaussian MSK (GMSK).

Why are there so many types of quadrature modulation? Recall that three parameters determine the performance of a modulation technique for RF applications: BER, spectral efficiency, and power efficiency. The above schemes exhibit different trade-offs among these parameters.

QPSK If in (3.52), the bit waveform is a rectangular pulse, a QPSK signal is obtained. From another point of view, in analogy with BPSK, one of four phases of a sinusoid is selected according to the symbol; that is, $x_{QPSK}(t) = \sqrt{2}A_c \cos(\omega_c t + k\pi/4)$, $k = 1, 3, 5, 7$.

Coherent detection of QPSK signals can be performed as shown in Fig. 3.34. Here, the circuit correlates the input with $\cos \omega_c t$ and $\sin \omega_c t$ to determine

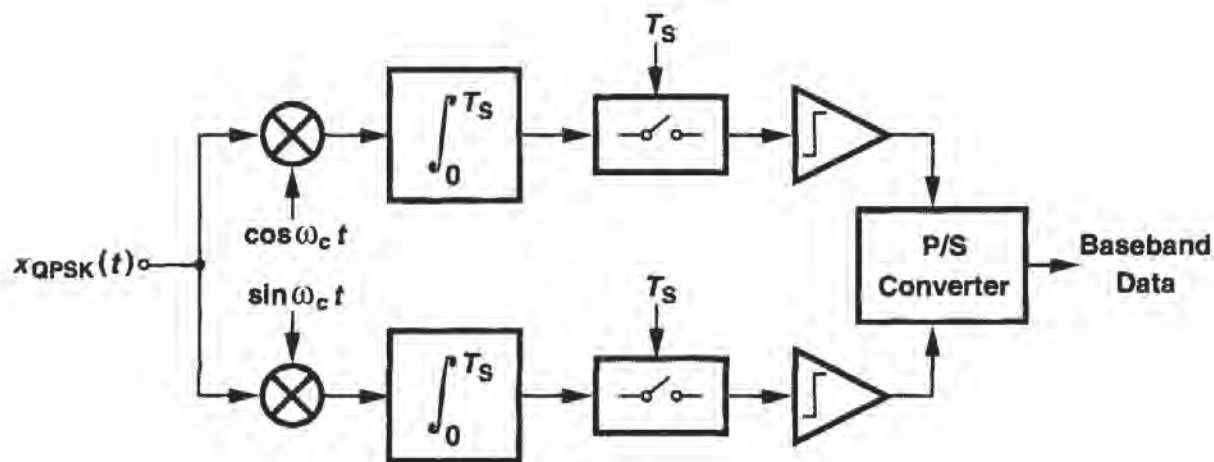


Figure 3.34 Coherent QPSK detection.

the most likely value for the symbol from the set $[\pm 1 \pm 1]$. A parallel-to-serial (P/S) converter then multiplexes every two bits in a symbol so as to produce the baseband binary data.

Upon comparison of Figs. 3.29(b) and 3.33, we may conclude that QPSK has a higher probability of error than does BPSK because each point in the constellation of Fig. 3.33 can be mistaken for *three* other points and the minimum distance between the points is the same in both schemes. A more careful analysis, however, indicates otherwise. First, since a fair comparison requires that the transmitters produce the same output power, we write the two modulated waveforms as $x_{BPSK}(t) = \pm A_c \cos \omega_c t$ and $x_{QPSK}(t) = A_c \cos(\omega_c t + k\pi/4)$, so that they both have an average power of $A_c^2/2$. Thus, $x_{QPSK}(t) = \pm(A_c/\sqrt{2}) \cos \omega_c t \pm (A_c/\sqrt{2}) \sin \omega_c t$, yielding the constellation in Fig. 3.35(b). Here, the points are closer to each other, further strengthening our postulate that QPSK has a higher BER. But a fair comparison also requires that the overall bit rates be equal in the two cases. This means that, for a baseband binary data with period T_b , the correlator in a BPSK receiver [Fig. 3.29(c)] integrates the input for T_b seconds, whereas those in a QPSK detector (Fig. 3.34) integrate the input for $T_s = 2T_b$ seconds. The sampled level (i.e., the energy) is therefore equal to $A_c^2 T_b/2$ in the first case and $(\sqrt{2}A_c/2)^2(2T_b) = A_c^2 T_b$ in the second case. Consequently, the *bit* energy in BPSK is less than the *symbol* energy in QPSK.

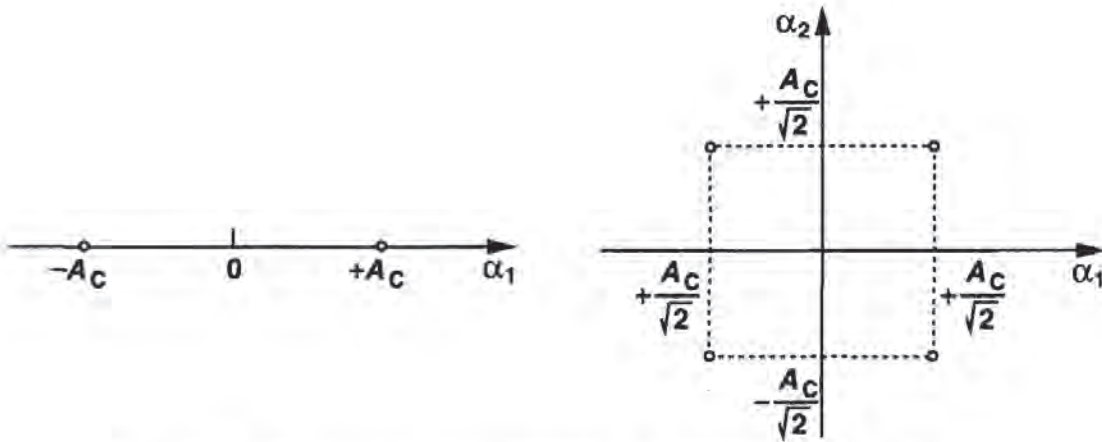


Figure 3.35 Signal constellation of BPSK and QPSK for equal output power.

We summarize the above observations as follows: compared with BPSK, QPSK has more adjacent points, smaller minimum distance, but higher symbol energy. If all of these are taken into account, it can be proved that BPSK and QPSK have nearly equal probabilities of error; that is, $P_e = Q(\sqrt{2E_b/N_0})$ [1].

Since a QPSK waveform can be viewed as the summation of two BPSK signals, the spectrum of QPSK waveforms is similar to that of their BPSK coun-

terparts but the bit stream is “slowed down” by a factor of two (S/P conversion) before quadrature multiplexing. Thus, the spectral density can be derived from that in (3.46) by substituting $2T_b$ for T_b .

An important drawback of QPSK is large phase changes at the end of each symbol. As depicted in Fig. 3.36, when the waveforms at the output of the S/P converter change simultaneously from, say, $[-1 \ -1]$ to $[+1 \ +1]$, the carrier experiences a 180° phase step or, equivalently, a transition between two diagonally opposite points in the constellation. As we will see in Section 3.4, such transitions are undesirable if the waveform is to be filtered and subsequently processed by a nonlinear power amplifier.

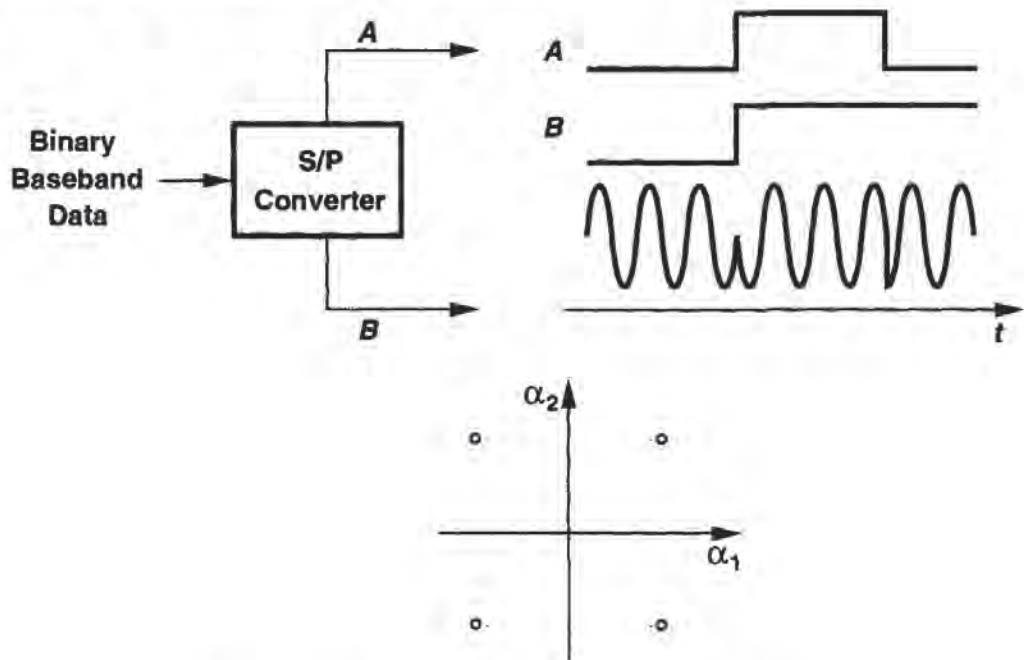


Figure 3.36 Phase transitions in a QPSK waveform.

A variant of QPSK that remedies the above drawback is offset QPSK. As shown in Fig. 3.37, the data streams are offset in time by half the symbol period after S/P conversion, thereby avoiding simultaneous transitions in waveforms at nodes A and B . The phase step is therefore only $\pm 90^\circ$. Fig. 3.38 illustrates the phase transitions in the time domain and in the constellation. The bit error rate and spectrum of OQPSK are identical to those of QPSK [1].

Despite the advantage of smaller phase change, OQPSK suffers from a critical drawback: it does not lend itself to “differential encoding” [4]. We defer this topic to Section 3.5, but point out here that differential encoding plays an important role in noncoherent receivers, the most popular type in today’s RF applications.

Another variant of QPSK is “ $\pi/4$ -QPSK” [3, 4]. In contrast to OQPSK, this scheme can be differentially encoded and is used in American and Japanese phone systems (Chapter 4). In $\pi/4$ -QPSK, the signal set consists of *two* QPSK

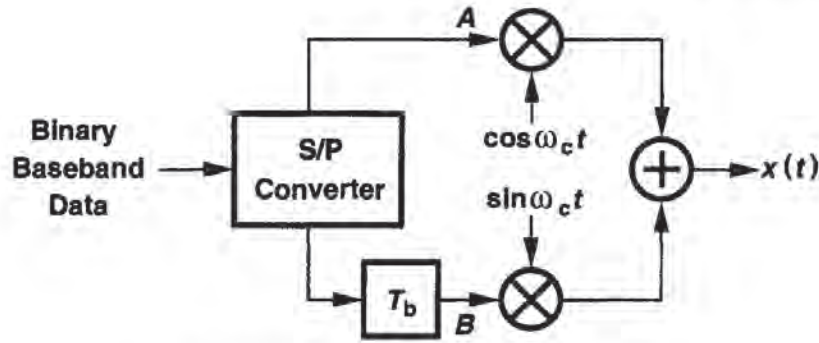


Figure 3.37 Offset QPSK modulator.

schemes, one rotated by 45° with respect to the other:

$$x_1(t) = A_c \cos\left(\omega_c t + k\frac{\pi}{4}\right) \quad k \text{ odd}, \quad (3.53)$$

$$x_2(t) = A_c \cos\left(\omega_c t + k\frac{\pi}{4}\right) \quad k \text{ even}. \quad (3.54)$$

As shown in Fig. 3.39, the modulation is performed by alternately taking the output from each QPSK generator.

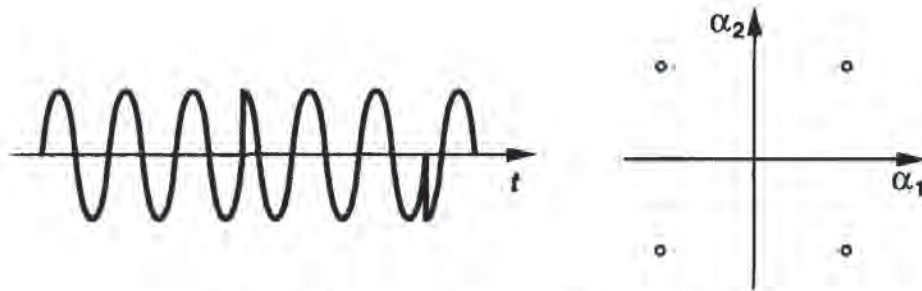
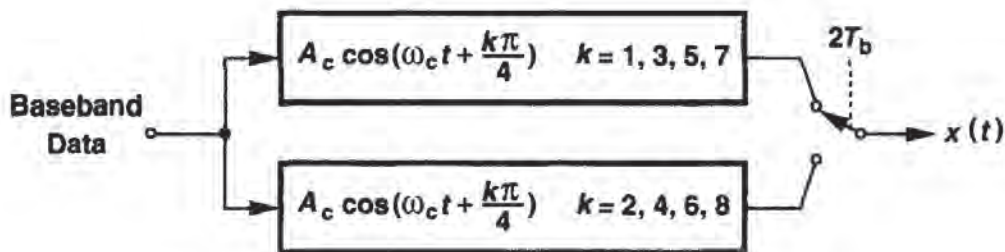
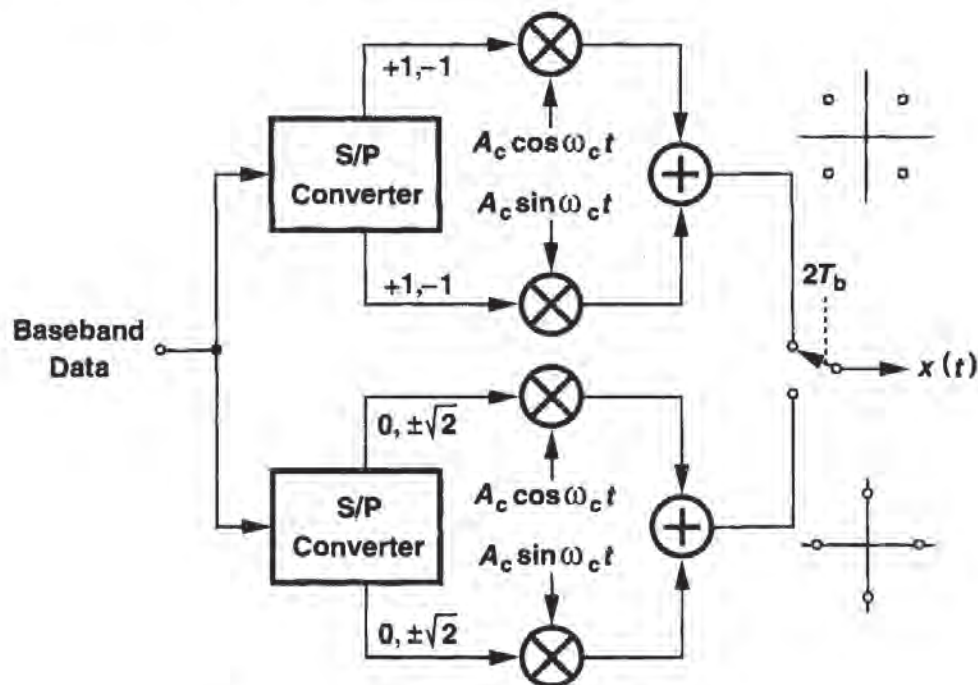


Figure 3.38 Phase transitions in OQPSK.

To better understand the operation, let us study the simple implementation of a $\pi/4$ -QPSK generator in Fig. 3.40. After S/P conversion, the digital signal levels are scaled and shifted so as to present ± 1 in the upper QPSK


 Figure 3.39 Conceptual generation of $\pi/4$ -QPSK signals.

Figure 3.40 Generation of $\pi/4$ -QPSK signals.

modulator and 0 and $\pm\sqrt{2}$ in the lower. The outputs are therefore equal to $x_1(t) = \alpha_1 \cos \omega_c t + \alpha_2 \sin \omega_c t$, where $[\alpha_1 \ \alpha_2] = [\pm A_c \ \pm A_c]$, and $x_2(t) = \beta_1 \cos \omega_c t + \beta_2 \sin \omega_c t$, where $[\beta_1 \ \beta_2] = [0 \ \pm\sqrt{2}A_c]$ and $[\pm\sqrt{2}A_c \ 0]$. Thus, the constellation alternates between the two depicted in Fig. 3.40. Now consider a baseband sequence of [11, 01, 10, 11, 01]. As shown in Fig. 3.41, the first pair, [1 1], is converted to $[+A_c \ +A_c]$ in the upper arm, producing $y(t) = A_c \cos(\omega_c t + \pi/4)$. The next pair, [0 1], is converted to $[0 \ -\sqrt{2}A_c]$ in the lower arm, yielding $y(t) = -\sqrt{2}A_c \cos \omega_c t$. Following the values of $y(t)$ for the entire sequence, we note that the points chosen from the two constellations appear as in Fig. 3.41(a) as a function of time. The key point here is that, since no two consecutive points are from the same constellation, the maximum phase step is 135° , 45° less than that in QPSK. This is illustrated in Fig. 3.41(b). Thus, in terms of the maximum phase change, $\pi/4$ -QPSK is an intermediate case between QPSK and OQPSK.

The spectrum and BER of $\pi/4$ -QPSK are identical to those of QPSK. We discuss the power efficiency of QPSK and its variants in Section 3.4.

In our treatment of the QPSK family, we have assumed that the baseband bits are represented by rectangular pulses. In band-limited applications, however, it is desirable to employ a pulse shape whose spectrum drops sharply in the adjacent channels. For example, the raised-cosine signal described in Section 2.2 provides both a tight spectrum and ISI-free operation.

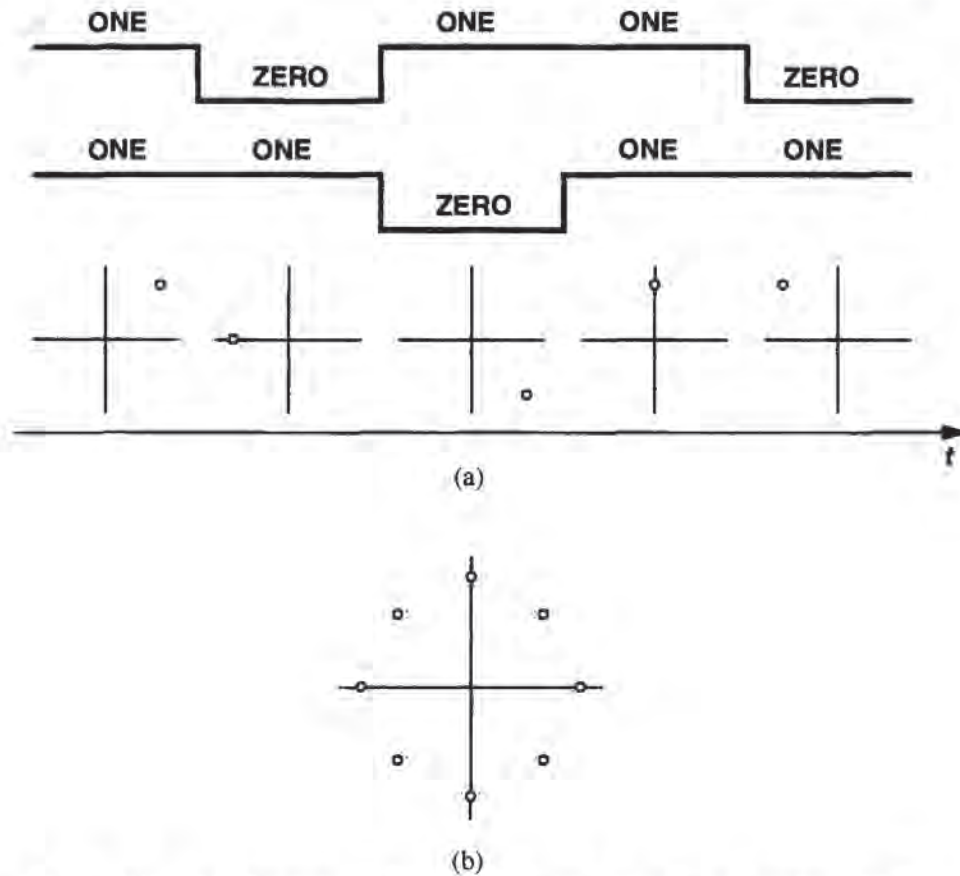


Figure 3.41 (a) Evolution of $\pi/4$ -QPSK in time domain, (b) possible phase transitions in the constellation.

In practice, the raised-cosine filter is decomposed into two sections, one placed in the transmitter and the other in the receiver (Fig. 3.42). Using a transfer function equal to the square root of (2.54), the two sections together allow Nyquist signaling while the section used in the receiver also operates as a matched filter. This technique is described in [5].

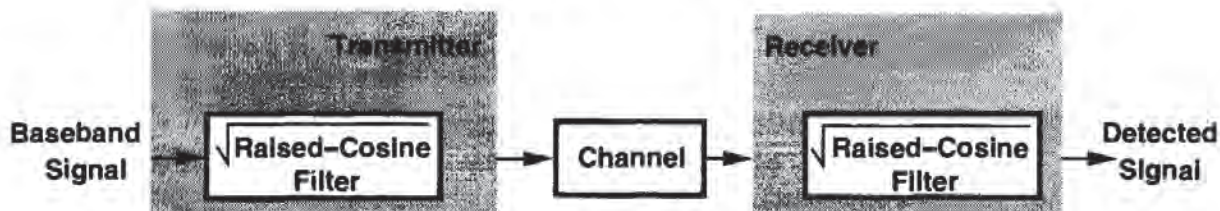


Figure 3.42 Decomposition of a raised-cosine filter into two sections.

MSK Minimum shift keying can be described from several different points of view, each providing more insight into the underlying principles. We first utilize OQPSK to arrive at MSK, but eventually view MSK as a type of FSK to understand GMSK.

Recall from Eq. (3.52) that in quadrature modulation, every two consecutive bits, b_m and b_{m+1} , are impressed on quadrature phases of a carrier. If b_m and b_{m+1} are rectangular pulses, abrupt phase changes arise in the modulated waveforms, leading to a wide spectrum and presenting difficulties in the design of power amplifiers (Section 3.4). This consideration warrants the use of “continuous phase modulation” schemes [6, 7]. One such scheme is MSK.

Suppose in OQPSK, we use half sinusoids, rather than rectangular pulses, to represent the levels that are multiplied by the carrier. More specifically, as shown in Fig. 3.43, let us multiply the levels in the upper arm by $\cos \omega_1 t$ and those in the lower arm by $\sin \omega_1 t$, where $\omega_1 = \pi/(2T_b)$ [9]. Thus, $x(t) = a_m \cos \omega_1 t \cos \omega_c t - a_{m+1} \sin \omega_1 t \sin \omega_c t$, where a_m and a_{m+1} are rectangular pulses toggling between $+1$ and -1 with a width of $2T_b$ and offset with respect to each other by T_b . As shown in Fig. 3.43, $x(t)$ exhibits no abrupt changes in the phase or, equivalently, in the slope of the amplitude. To understand why, first note that a_m changes only at $t = (2k+1)T_b$ and a_{m+1} only at $t = 2kT_b$. Now suppose a_m goes from $+1$ to -1 while $a_{m+1} = +1$. Thus, $x(t)$ goes from $2 \cos(\omega_1 + \omega_c)t$ to $2 \cos[(\omega_c - \omega_1)t + \pi]$, where $t = (2k+1)T_b$. As a result, the phase change is $\Delta\phi = (\omega_c - \omega_1)t + \pi - (\omega_1 + \omega_c)t = -2\omega_1 t + \pi = -2k\pi$ because $\omega_1 = \pi/(2T_b)$. The same can be proved for the other three possible transitions in a_m and a_{m+1} , indicating that the phase is always continuous.

While exhibiting the same error rate as QPSK, MSK exhibits a sharper decay in its spectrum than the rectangular-pulse QPSK family. The smooth phase transitions in MSK lower the signal power in the sidelobes of the spectrum, but at the cost of widening the main lobe. As a rule of thumb, if the phase of the bandpass signal has finite derivatives up to and including order n , then the spectrum of the signal decays at a rate proportional to $f^{2(n+1)}$ [8]. In fact, for MSK, we have

$$S_{MSK}(f) = \frac{16A_c^2 T_b}{\pi^2} \left\{ \frac{\cos^2[2\pi T_b(f - f_c)]}{[1 - 16T_b^2(f - f_c)^2]^2} + \frac{\cos^2[2\pi T_b(f + f_c)]}{[1 - 16T_b^2(f + f_c)^2]^2} \right\}, \quad (3.55)$$

that is, a decay proportional to f^4 . This should be compared to (3.46).

The above rule of thumb suggests that even more spectrally efficient schemes can be obtained if the phase change is made smoother than that in MSK. This has led to a more general class of MSK signals [8], of which Gaussian MSK is an example. To arrive at GMSK, we first state without proof that the MSK signal of Fig. 3.43 can also be written as

$$x_{MSK}(t) = \sqrt{2}A_c \cos \left[\omega_c t + \int_{-\infty}^t \sum_m b_m p(t - mT_b) dt \right], \quad (3.56)$$

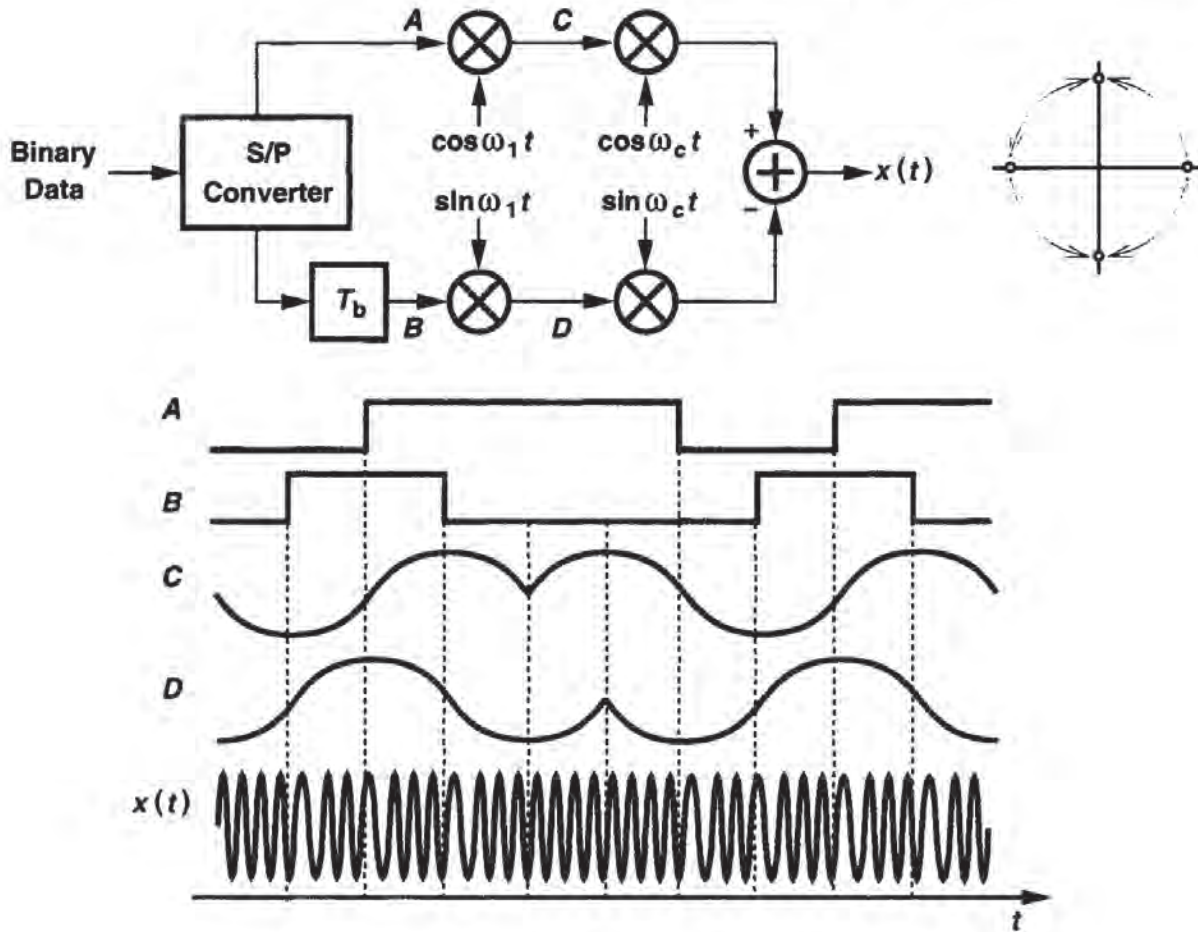


Figure 3.43 MSK modulation and signal constellation.

where the summation represents the binary baseband signal; that is, $b_m = \pm 1$ and $p(t)$ is a rectangular pulse of width T_b . In GMSK, on the other hand, $p(t)$ is a smoother function. Note that (3.56) can be viewed as frequency modulation and, with a rectangular $p(t)$, as FSK.

In GMSK, the pulse shape $p(t)$ is obtained by passing the baseband rectangular pulses through a filter with a Gaussian impulse response, $h(t) = \exp(-\alpha t^2)$. As depicted in Fig. 3.44, to obtain the output of the filter, $x_{BB}(t) * h(t)$, we can calculate $(dx_{BB}/dt) * h(t)$ and integrate the result. This method shows that for a narrow Gaussian response (large α), the filter output is close to a rectangular pulse with width T_b , and as $h(t)$ becomes wider, so does the output.

The derivation of the GMSK power spectral density is difficult, but we intuitively expect the spectrum to decay faster if the time-domain modulating pulse, $p(t)$, becomes smoother. For example, if $\alpha = 1$, the 99% bandwidth is less than $1.2/T_b$ [10].

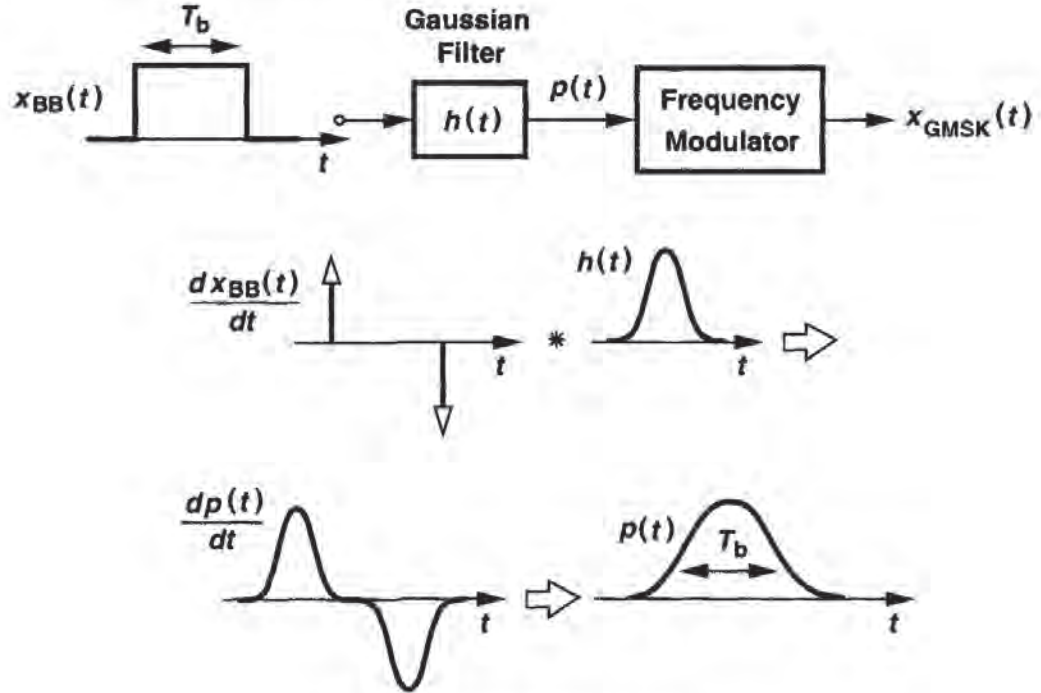


Figure 3.44 GMSK modulation.

While our analysis indicates that the lower the value of α , the narrower the spectrum, the choice of α depends on another factor as well: intersymbol interference. In contrast to previous modulation schemes, GMSK allows the modulating pulse to be wider than T_b , thus suffering from ISI. As α decreases, ISI becomes more significant. In typical systems, α is in the vicinity of 0.3, requiring ISI mitigation in the receiver. The power efficiency of MSK and GMSK is described in Section 3.4.

The spectra and error probabilities of various modulation schemes are given in [11, 12, 13].

3.4 POWER EFFICIENCY OF MODULATION SCHEMES

In our study of various modulation techniques thus far, we have frequently mentioned that power efficiency is an important aspect. In this section, we deal with this issue.

3.4.1 Constant- and Variable-Envelope Signals

A modulated waveform $x(t) = A(t) \cos[\omega_c t + \phi(t)]$ is said to have a constant envelope if $A(t)$ does not vary with time. Otherwise, we say the signal has a variable envelope.

Constant- and variable-envelope signals behave differently in a nonlinear system. Suppose $A(t) = A_c$ and the system exhibits a third-order memoryless nonlinearity:

$$y(t) = \alpha_3 x^3(t) + \dots \quad (3.57)$$

$$= \alpha_3 A_c^3 \cos^3[\omega_c t + \phi(t)] + \dots \quad (3.58)$$

$$= \frac{\alpha_3 A_c^3}{4} \cos[3\omega_c t + 3\phi(t)] + \frac{3\alpha_3 A_c^3}{4} \cos[\omega_c t + \phi(t)]. \quad (3.59)$$

The first term in (3.59) represents a modulated signal around $\omega = 3\omega_c$. Since the bandwidth of the original signal, $A_c \cos[\omega_c t + \phi(t)]$, is typically much less than ω_c , we note from Carson's rule that the bandwidth occupied by $\cos[3\omega_c t + 3\phi(t)]$ is also quite small. Thus, the shape of the spectrum in the vicinity of ω_c remains unchanged.

Now consider a variable-envelope signal applied to the above nonlinear system. Writing $x(t)$ as

$$x(t) = x_I(t) \cos \omega_c t - x_Q(t) \sin \omega_c t, \quad (3.60)$$

where $x_I(t)$ and $x_Q(t)$ are the baseband I and Q components, we have

$$y(t) = \alpha_3 [x_I(t) \cos \omega_c t - x_Q(t) \sin \omega_c t]^3 + \dots \quad (3.61)$$

$$= \alpha_3 x_I^3(t) \frac{\cos 3\omega_c t + 3 \cos \omega_c t}{4} - \alpha_3 x_Q^3(t) \frac{-\sin 3\omega_c t + 3 \sin \omega_c t}{4} + \dots, \quad (3.62)$$

Thus, the output contains the spectra of $x_I^3(t)$ and $x_Q^3(t)$ centered around ω_c . Since these components generally exhibit a broader spectrum than do $x_I(t)$ and $x_Q(t)$, we say the spectrum "grows" when a variable-envelope signal passes through a nonlinear system.

An exception to the above observation occurs if the baseband pulse shape is rectangular. As higher powers of a rectangular waveform are still rectangular, the modulated signal spectrum has a $\text{sinc}^2 \omega T_b$ shape for any nonlinearity.

3.4.2 Spectral Regrowth

Consider the simple transmitter illustrated in Fig. 3.45. Here, a low-pass filter precedes the modulator to limit the bandwidth of the signal, thereby suppressing spectral leakage into adjacent channels. The PA amplifies and buffers the result, delivering the required signal level to the antenna.

What is the effect of filtering on a digitally modulated signal? Intuitively, we expect limiting the bandwidth tends to smooth out the abrupt transitions in the time domain. For example, a QPSK waveform appears as shown in Fig. 3.45 after filtering, exhibiting greater variation in its *envelope* as the filter bandwidth

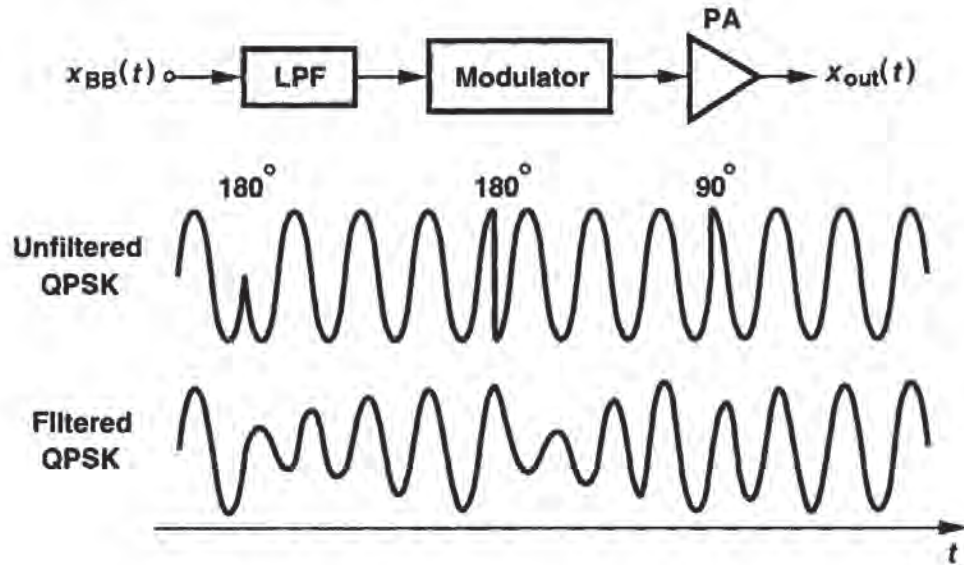


Figure 3.45 Effect of bandlimiting on the envelope of a QPSK waveform.

decreases. We also note that envelope variation due to filtering becomes more pronounced at greater phase steps.

What is the implication of envelope variation? The one-to-one correspondence between the band-limited spectrum and the time-varying envelope indicates that if the power amplifier is to maintain the spectrum to the limited bandwidth, then it must also amplify the envelope variations *linearly*. This can be seen by expressing the modulated waveform as $x(t) = x_I(t) \cos \omega_c t - x_Q(t) \sin \omega_c t$. If the PA exhibits significant nonlinearity, then the shape of $x_I(t)$ and $x_Q(t)$ is not preserved and the spectrum is not limited to the desired bandwidth. This effect is called “spectral regrowth” [14, 15] and can be quantified by the relative adjacent channel power [15].

The above discussion can be summarized as follows. Abrupt phase changes in a digitally modulated waveform, for example, QPSK, result in envelope variations if a filter limits the bandwidth. Such variations in turn require a linear power amplifier so as to avoid spectral regrowth. As explained in Chapter 9, linear PAs are typically less efficient than nonlinear PAs. For a 40% efficiency—which is rarely achieved in linear PAs—and a 1-W output power, the PA consumes a total of 2.5 W, a considerable amount with respect to the power dissipation of the rest of a portable phone. Nonlinear PAs, on the other hand, exhibit efficiencies as high as 60%. Thus, it is desirable to employ modulation schemes that do not experience spectral regrowth when processed by nonlinear amplifiers. This aspect of modulation formats is called power efficiency.

A simple example of power-efficient modulation is FM. From $x_{FM}(t) = A_c \cos[\omega_c t + m \int x_{BB}(t) dt]$, we see that FM waveforms have no abrupt phase change and exhibit a constant envelope. FM and FSK signals can therefore be amplified by means of nonlinear PAs with no spectral regrowth.

From the illustration of Fig. 3.45, we conclude that filtered QPSK is quite susceptible to spectral regrowth. For example, if root raised-cosine filtering is used in the baseband so as to achieve a narrowband, the resulting envelope variation requires a relatively linear power amplifier [17, 16]. This drawback is less severe in OQPSK and $\pi/4$ -QPSK by virtue of smaller phase steps allowed in these schemes.

More power efficient than the QPSK family are MSK signals because their constant envelope allows processing by nonlinear PAs. However, MSK signals typically require a wider band than QPSK waveforms with raised-cosine filtering.

In summary, digital modulation schemes exhibit a trade-off between spectral efficiency and power efficiency. As the signal bandwidth is more limited—by filtering or pulse shaping—the power amplifier must achieve a higher linearity so as to avoid spectral regrowth. Different wireless systems have aimed at one of these extremes: pulse-shaped QPSK modulation with narrow channels but linear PAs or MSK modulation with wider channels but nonlinear PAs (Chapter 4).

3.5 NONCOHERENT DETECTION

In Section 3.3.1, we saw that matched filters, an example of which are coherent detectors (correlators), provide the highest signal-to-noise ratio and hence the lowest probability of error. Coherent detectors, however, require that the phase of the local oscillator in the receiver bear a strict relation to that of the received signal. Called “carrier recovery,” establishing such a phase alignment necessitates substantial circuit complexity and becomes especially difficult at low signal levels in the presence of interferers and signal fading. For this reason, many RF systems employ noncoherent detection despite its somewhat inferior performance.

Noncoherent FSK Detection In Section 3.3.1, we briefly studied a noncoherent FSK detector. Shown in Fig. 3.46, this configuration employs two bandpass filters (for binary modulation), each followed by an envelope detector, to determine which frequency is received. Since $\omega_1 - \omega_2$ is typically much less than ω_1 and ω_2 , the input spectrum is first translated to a lower frequency so as to relax the Q and the precision of the center frequency required of the bandpass filters.

Derivation of the probability of error for this detector is rather involved. We only state the result for the practical case of sufficient difference between ω_1 and ω_2 so that their spectral overlap is negligible [1]

$$P_e = \frac{1}{2} \exp \frac{-E_b}{2T_b B_p N_0}, \quad (3.63)$$

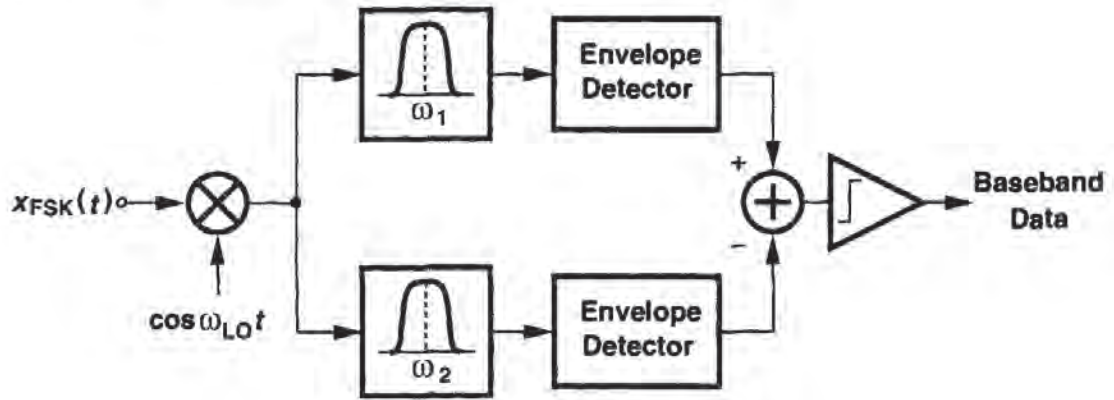


Figure 3.46 Noncoherent FSK detection.

where B_p is the (effective) bandwidth of each bandpass filter and E_b is the bit energy, $E_b = A_c^2 T_b / 2$. If $B_p = 1/T_b$, then

$$P_e = \frac{1}{2} \exp \frac{-E_b}{2N_0}. \quad (3.64)$$

For error rates on the order of 10^{-3} , (3.64) requires an E_b/N_0 that is 1.5 dB greater than that in coherent FSK detection [Eq. (3.51)]. In many applications, the lower complexity justifies this loss in the performance.

Differential Phase Shift Keying While in FSK it is possible to determine the frequency corresponding to each bit (using bandpass filters or frequency demodulators), in PSK the phase relates to the time origin and has no “absolute” meaning. For example, a 90° phase shift in a QPSK waveform converts the constellation to a similar one, but with all the symbols interpreted incorrectly. Thus, simple PSK waveforms cannot be detected noncoherently. However, if the information lies in the phase *change* from one bit (or symbol) to the next, then a time origin is not required and noncoherent detection is possible. This is accomplished through “differential” encoding and decoding of the baseband signal before modulation and after demodulation, respectively.

Let us consider binary differential PSK (DPSK). The rule for differential encoding is that if the present input bit is a ONE, then the output state of the encoder does not change, and vice versa. This requires an extra starting bit (of arbitrary value). The concept can be better understood by considering the implementation depicted in Fig. 3.47(a). An exclusive-NOR (XNOR) gate compares the present output bit, $D_{\text{out}}(mT_b)$, with the present input bit, $D_{\text{in}}(mT_b)$, to determine the next output state,

$$D_{\text{out}}[(m+1)T_b] = \overline{D_{\text{in}}(mT_b) \oplus D_{\text{out}}(mT_b)}, \quad (3.65)$$

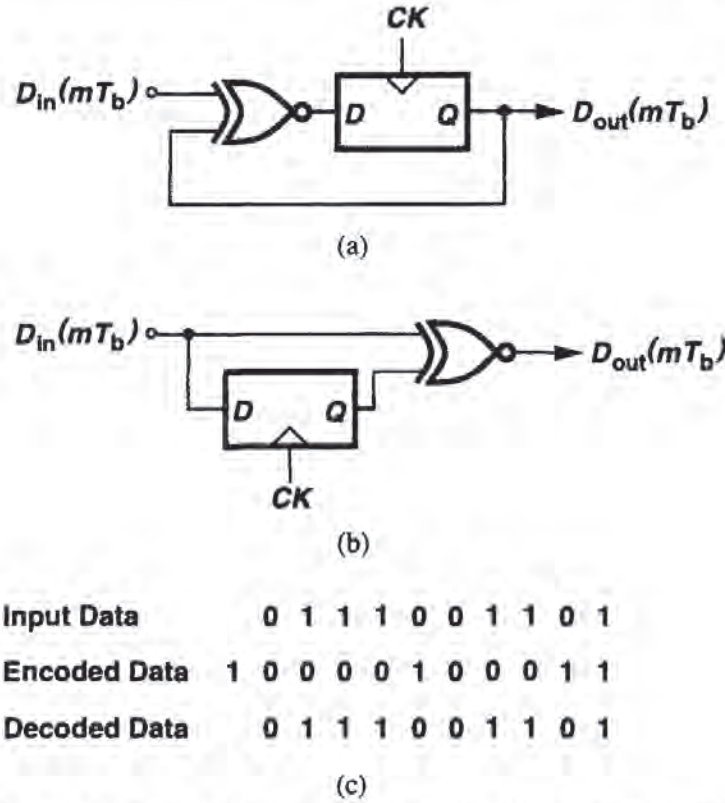


Figure 3.47 (a) Differential encoding, (b) differential decoding, (c) example of encoded and decoded sequence.

implying that if $D_{in}(mT_b) = 1$, then $D_{out}[(m+1)T_b] = D_{out}(mT_b)$, and if $D_{in}(mT_b) = 0$ then $D_{out}[(m+1)T_b] = \overline{D_{out}(mT_b)}$. The extra starting bit mentioned above corresponds to the state of the flipflop before the data sequence begins.

The DPSK signal received by a system can be decoded by a reverse operation, depicted in Fig. 3.47(b). It can be easily shown that phase shifts in the encoded signal do not alter the differentially decoded output of the receiver.

Differential signaling does not provide the lowest bit error rate because it is not based on matched filtering. From another point of view, in coherent demodulation the signal is compared with a low-noise reference (oscillator), whereas in differential detection two noisy signals are compared with each other [12]. It can be shown that the probability of error for DPSK is

$$P_e = \frac{1}{2} \exp \frac{-E_b}{N_0}, \quad (3.66)$$

where $E_b = A_c^2 T_b / 2$. This result indicates that for an error rate of less than 10^{-3} , DPSK requires approximately 3 dB higher E_b/N_0 than does coherent PSK [Eq. (3.43)].

REFERENCES

- [1] L. W. Couch, *Digital and Analog Communication Systems*, 4th ed., New York: Macmillan, 1993.
- [2] H. E. Rowe, *Signals and Noise in Communication Systems*, Princeton, NJ: Van Nostrand, 1965.
- [3] P. A. Baker, "Phase-Modulated Data Sets for Serial Transmission at 2000 and 2400 Bits per Second," Part I, *AIEE Trans. Communication Electronics*, pp. 166–171, July 1962.
- [4] Y. Akaiwa and Y. Nagata, "Highly Efficient Digital Mobile Communication with a Linear Modulation Method," *IEEE J. of Selected Areas in Communications*, Vol. 5, pp. 890–895, June 1987.
- [5] R. D. Gitlin, J. F. Hayes, and S. B. Weinstein, *Data Communications Principles*, New York: Plenum Press, 1992.
- [6] C. E. W. Sundberg, "Continuous Phase Modulation," *IEEE Comm. Mag.*, Vol. 24, pp. 25–38, April 1986.
- [7] J. B. Anderson and C. E. W. Sundberg, "Advances in Constant Envelope Coded Modulation," *IEEE Comm. Mag.*, Vol. 29, pp. 36–56, December 1991.
- [8] R. E. Ziemer and R. L. Peterson, *Digital Communication and Spread Spectrum Systems*, New York: Macmillan, 1985.
- [9] S. Pasupathy, "Minimum Shift Keying: A Spectrally Efficient Modulation," *IEEE Comm. Mag.*, Vol. 19, pp. 14–22, July 1979.
- [10] K. Murota and K. Hirade, "GMSK Modulation for Digital Mobile Radio Telephony," *IEEE Trans. Comm.*, Vol. 29, pp. 1044–1051, July 1981.
- [11] T. S. Rappaport, *Wireless Communications, Principles and Practice*, Englewood Cliffs, NJ: Prentice Hall, 1996.
- [12] B. Sklar, *Digital Communications*, Englewood Cliffs, NJ: Prentice Hall, 1988.
- [13] K. Feher, *Wireless Digital Communications*, Englewood Cliffs, NJ: Prentice Hall, 1995.
- [14] D. H. Morais and K. Feher, "The Effects of Filtering and Limiting on the Performance of QPSK, Offset QPSK, and MSK Signals," *IEEE Trans. Communications*, Vol. 28, pp. 1999–2009, December 1980.

- [15] J. F. Sevic and J. Staudinger, "Simulation of Adjacent Channel Power for Digital Wireless Communication Systems," *Microwave J.*, pp. 66–80, October 1996.
- [16] S. Ariyavisitakul and T. P. Liu, "Characterizing the Effects of Nonlinear Amplifiers on Linear Modulation for Digital Portable Radio Communications," *IEEE Trans. Veh. Tech.*, Vol. 39, pp. 383–389, November 1990.
- [17] J. Boccuzzi, "Performance Evaluation of Nonlinear Transmit Power Amplifiers for North American Digital Cellular Portables," *IEEE Trans. Veh. Tech.*, Vol. 44, pp. 220–228, May 1995.

MULTIPLE ACCESS TECHNIQUES AND WIRELESS STANDARDS

The modulation and detection techniques described in Chapter 3 provide the basis for communication between only one transmitter and one receiver. For a large number of transceivers in a network, additional methods are required to ensure proper communication among multiple users. Such methods are called “multiple access techniques.”

The complexity of wireless systems mandates that they conform to a “standard.” In addition to modulation and multiple access techniques, a wireless standard includes such details as frequency bands, timing, and data coding, while defining precise tests for the performance of transceivers.

Following a brief overview of the properties of mobile communication systems, we describe in this chapter three general multiple access techniques that are commonly used in RF systems. These are frequency-division multiple access (FDMA), time-division multiple access (TDMA), and code-division multiple access (CDMA). We then study several popular wireless standards, namely, North American Digital Communication, Global System for Mobile Communications, Qualcomm CDMA, and Digital European Cordless Telephone.

4.1 MOBILE RF COMMUNICATIONS

A mobile system is one in which users can physically move while communicating with one another. Examples include pagers, cellular phones, and cordless phones. It is the mobility that has made RF communications powerful and popular. The transceiver carried by the user is called the “mobile unit” (or simply the “mobile”), the “terminal,” or the “hand-held unit.” The complexity of the wireless infrastructure often demands that the mobiles communicate only through a fixed, relatively expensive unit called the “base station.” Each mobile receives and transmits information from and to the base station via two RF

channels called the “forward channel” or “downlink” and the “reverse channel” or “uplink,” respectively. Most of our treatment in this book relates to the mobile unit because, compared to the base station, hand-held units constitute a much larger portion of the market and their design is much more similar to other types of RF systems.

Cellular System With the limited available spectrum (e.g., 25 MHz around 900 MHz), how do hundreds of thousands of people communicate in a crowded metropolitan area? To answer this question, we first consider a simpler case: thousands of FM radio broadcasting stations may operate in a country in the 88–108 MHz band. This is possible because stations that are physically far enough from each other can use the same carrier frequency (“frequency reuse”) with negligible mutual interference (except at some point in the middle where the stations are received with comparable signal levels). The minimum distance between two stations that can employ equal carrier frequencies depends on the signal power produced by each.

In mobile communications, the concept of frequency reuse is implemented in a “cellular” structure, where each cell is configured as a hexagon and surrounded by six other cells [Fig. 4.1(a)]. The idea is that if the center cell uses a frequency f_1 for communication, the six neighboring cells cannot utilize this frequency but the cells beyond the immediate neighbors may. In practice, more efficient frequency assignment leads to the “7-cell” reuse pattern shown in Fig. 4.1(b). Note that in reality each cell utilizes a group of frequencies.

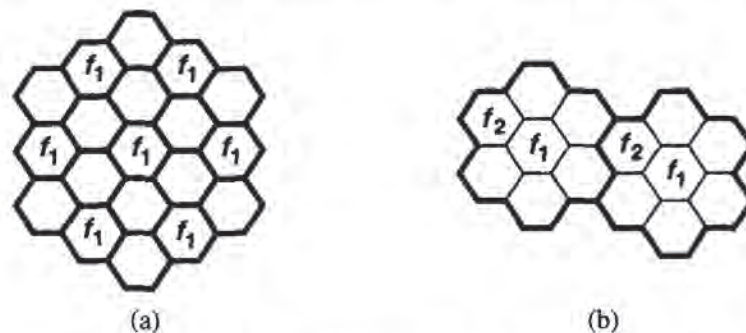


Figure 4.1 (a) Simple cellular system, (b) 7-cell reuse pattern.

The mobile units in each cell of Fig. 4.1(b) are served by a base station, and all of the base stations are controlled by a “mobile telephone switching office” (MTSO).

Co-Channel Interference An important issue in a cellular system is how much two cells that use the same frequency interfere with each other (Fig. 4.2). Called co-channel interference (CCI), this effect depends on the *ratio* of the distance between two co-channel cells to the cell radius and is independent of the transmitted power. Given by the frequency reuse plan, this ratio is

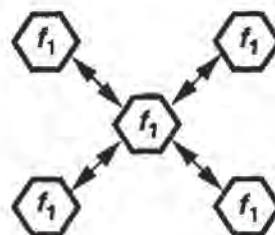


Figure 4.2 Co-channel interference.

approximately equal to 4.6 for the 7-cell pattern of Fig. 4.1(b) [1]. It can be shown that this value yields a signal-to-co-channel interference ratio of 18 dB [1].

Handoff What happens when a mobile unit “roams” from cell A to cell B (Fig. 4.3)? Since the power level received from the base station in cell A is insufficient to maintain communication, the mobile must change its server to the base station in cell B. Furthermore, since adjacent cells do not use the same group of frequencies, the channel must also change. Called “handoff,” this process is performed by the MTSO. Once the level received by the base station in cell A drops below a threshold, the MTSO hands off the mobile to the base station in cell B, hoping that the latter is close enough. This strategy fails with relatively high probability, resulting in dropped calls.

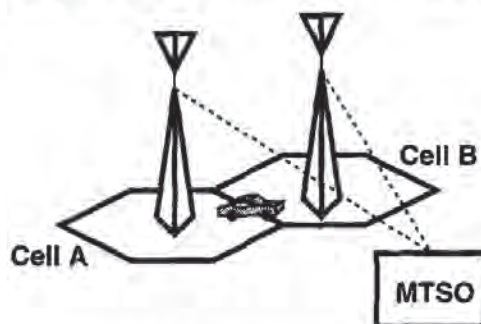


Figure 4.3 Problem of handoff.

To improve the handoff process, second-generation cellular systems allow the mobile unit to measure the received signal level from different base stations, thus performing handoff when the path to the second base station has sufficiently low loss [1].

Path Loss and Multipath Fading Propagation of signals in a mobile communication environment is quite complex. We briefly describe some of the important concepts here. Signals propagating through free space experience a power loss proportional to the square of the distance, d , from the source. In reality, however, the signal travels through both a direct path and an indirect, reflective path (Fig. 4.4). It can be shown that in this case, the loss increases with the *fourth* power of the distance [1]. In crowded areas, the actual loss profile may be proportional to d^2 for some distance and d^4 for another.

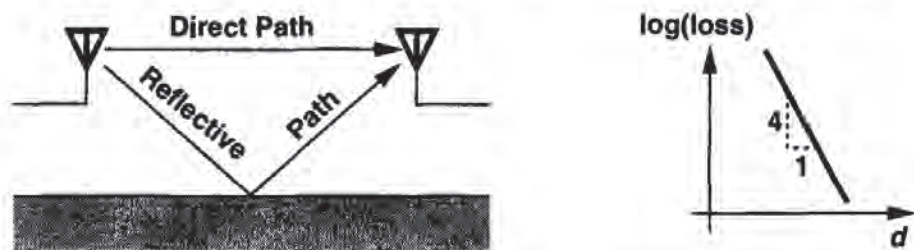


Figure 4.4 Indirect signal propagation and resulting loss profile.

In addition to the overall loss profile depicted in Fig. 4.4, another mechanism gives rise to fluctuations in the received signal level as a function of distance. Since the two signals shown in Fig. 4.4 generally experience different phase shifts, possibly arriving at the receiver with opposite phases and roughly equal amplitudes, the net received signal may be very small. Called “multipath fading” and illustrated in Fig. 4.5, this phenomenon introduces enormous variations in the signal level as the receiver moves by a fraction of the wavelength.

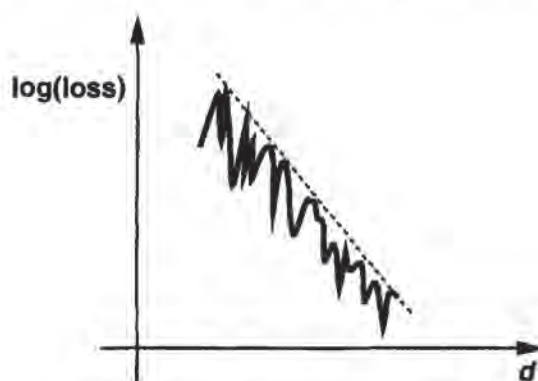


Figure 4.5 Multipath loss profile.

In reality, since the transmitted signal is reflected by many buildings and moving cars, the fluctuations are quite irregular. Nonetheless, the overall received signal can be expressed as

$$x_R(t) = a_1(t) \cos(\omega_c t + \theta_1) + a_2(t) \cos(\omega_c t + \theta_2) + \cdots + a_n \cos(\omega_c t + \theta_n) \quad (4.1)$$

$$= \left[\sum_{j=1}^n a_j(t) \cos \theta_j \right] \cos \omega_c t - \left[\sum_{j=1}^n a_j(t) \sin \theta_j \right] \sin \omega_c t. \quad (4.2)$$

Note that for large n , each summation has a Gaussian distribution. Denoting the first summation by A and the second by B , we have

$$x_R(t) = \sqrt{A^2 + B^2} \cos(\omega_c t + \phi), \quad (4.3)$$

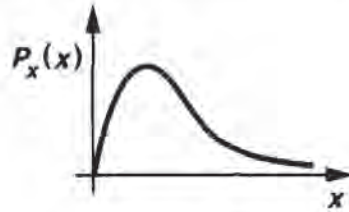


Figure 4.6 Rayleigh distribution.

where $\phi = \tan^{-1}(B/A)$. It can be shown that the amplitude, $A_m = \sqrt{A^2 + B^2}$, has a Rayleigh distribution (Fig. 4.6) [2], exhibiting losses greater than 10 dB below the mean for approximately 6% of the time.

From the above discussion, we conclude that in an RF system the transmitter output power and the receiver dynamic range must be chosen so as to accommodate signal level variations due to both the overall path loss (roughly proportional to d^4) and the deep multipath fading effects. While it is theoretically possible that multipath fading yields a zero amplitude (infinite loss) at a given distance, the probability of this event is negligible because moving objects in a mobile environment tend to “soften” the fading.

Diversity The effect of fading can be lowered by adding redundancy to the transmission or reception of the signal. “Space diversity” or “antenna diversity” employs two or more antennas spaced apart by a significant fraction of the wavelength so as to achieve a higher probability of receiving a nonfaded signal.

“Frequency diversity” refers to the case where multiple carrier frequencies are used, with the idea that fading is unlikely to occur simultaneously at two frequencies sufficiently far from each other.

“Time diversity” is another technique whereby the data is transmitted or received more than once to overcome short-term fading.

Delay Spread Suppose two signals in a multipath environment experience roughly equal attenuation but different delays. This is possible because the absorption coefficient and phase shift of reflective or refractive materials vary widely, making it likely for two paths to exhibit equal loss and unequal delays. Addition of two such signals yields $x(t) = A \cos \omega(t - \tau_1) + A \cos \omega(t - \tau_2) = 2A \cos[(2\omega t - \omega\tau_1 - \omega\tau_2)/2] \cos[\omega(\tau_1 - \tau_2)/2]$, where the second cosine factor relates the fading to the “delay spread,” $\Delta\tau = \tau_1 - \tau_2$. An important issue here is the frequency dependence of the fade. As illustrated in Fig. 4.7, small delay spreads yield a relatively flat fade, whereas large delay spreads introduce considerable variation in the spectrum.

In a multipath environment, many signals arrive at the receiver with different delays, yielding rms delay spreads as large as several microseconds and hence fading bandwidths of several hundreds of kilohertz. Thus, an entire communication channel may be suppressed during such a fade.

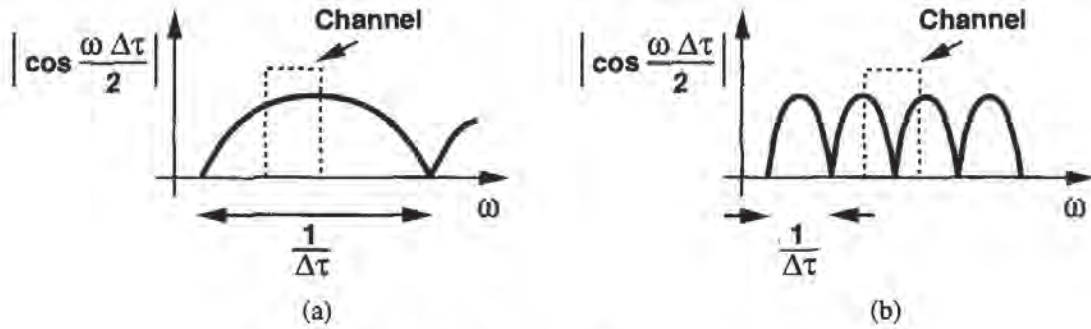


Figure 4.7 (a) Flat and (b) frequency-selective fading.

Large delay spreads introduce another difficulty as well: if the delay spread is comparable with the bit period of the digital modulating waveform, then multiple replicas of the signal are received with different delays, giving rise to considerable intersymbol interference.

Interleaving The nature of multipath fading and the signal processing techniques used to alleviate this issue is such that errors occur in clusters of bits. In order to lower the effect of these errors, the baseband bit stream in the transmitter undergoes “interleaving” before modulation. An interleaver in essence scrambles the time order of the bits according to an algorithm known by the receiver [1]. Interleaving can also be viewed as a type of time diversity with no overhead (although it entails some latency).

4.2 MULTIPLE ACCESS TECHNIQUES

4.2.1 Time- and Frequency-Division Duplexing

The simplest case of multiple access is the problem of two-way communication by a transceiver, a function called “duplexing.” In old walkie-talkies, for example, the user would press the “talk” button to transmit while disabling the receive path and release the button to listen while disabling the transmit path. This can be considered a simple form of “time-division duplexing” (TDD), whereby the same frequency band is utilized for both transmit (TX) and receive (RX) paths, but the system transmits for half of the time and receives for the other half. Illustrated in Fig. 4.8, TDD is usually performed fast enough to be transparent to the user.

Another approach to duplexing is to employ two different frequency bands for the transmit and receive paths. Called “frequency-division duplexing” (FDD) and shown in Fig. 4.9, this technique incorporates bandpass filters to isolate the two paths, allowing simultaneous transmission and reception. Since two such transceivers cannot communicate directly, the TX band must be translated to the RX band at some point. In wireless networks, this translation is performed in the base station.

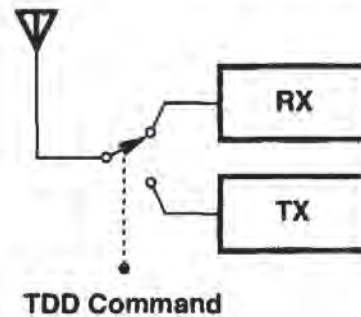


Figure 4.8 Time-division duplexing.

It is instructive to contrast the two duplexing methods by considering their merits and drawbacks. In TDD, an RF switch with a loss less than 1 dB follows the antenna to alternately enable and disable the TX and RX paths. Even though the transmitter output power may be 100 dB above the receiver input signal, the two paths do not interfere because the transmitter is disabled during reception. Furthermore, TDD allows direct (“peer-to-peer”) communication between two transceivers, an especially useful feature in short-range, local area network applications. The primary drawback of TDD is that the strong signals generated by all of the nearby mobile transmitters fall in the receive band, thus desensitizing the receiver.

In FDD systems, the two front-end bandpass filters are combined to form a “duplexer filter.” While making the receivers immune to the strong signals transmitted by other mobile units, FDD suffers from a number of issues. First, components of the transmitted signal that leak into the receive band are attenuated by typically only about 50 dB. Second, owing to the trade-off between the loss and the quality factor of filters, the loss of the duplexer is typically quite higher than that of a TDD switch. Note that a loss of, say, 3 dB in the RX path of the duplexer raises the overall noise figure by 3 dB (Chapter 2) and the same loss in the TX path of the filter means that only 50% of the signal power reaches the antenna. While providing acceptable isolation between the TX and RX paths, typical duplexers indeed exhibit a loss of 2 to 3 dB.

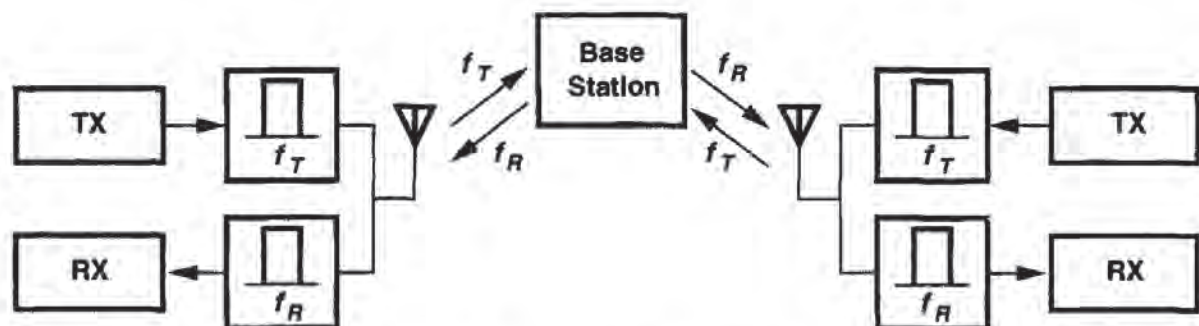


Figure 4.9 Frequency-division duplexing.

Another issue in FDD is the spectral leakage to adjacent channels in the transmitter output. This occurs when the power amplifier is turned on and off to save energy or when the local oscillator driving the modulator undergoes a transient. By contrast, in TDD such transients can be timed to end before the antenna is switched to the power amplifier output.

Despite the above drawbacks, FDD is employed in many RF systems, particularly in cellular communications, because it isolates the receivers from the signals produced by mobile transmitters.

4.2.2 Frequency-Division Multiple Access

In order to allow simultaneous communication among multiple transceivers, the available frequency band can be partitioned into many channels, each of which is assigned to one user. Called frequency-division multiple access (Fig. 4.10), this technique should be familiar within the context of radio and television broadcasting, where the channel assignment does not change with time. In multiple-user, two-way communications, on the other hand, the channel assignment may remain fixed only until the end of the call; after the user hangs up the phone, the channel becomes available to others. Note that in FDMA with FDD, two channels are assigned to each user, one for transmit and another for receive.

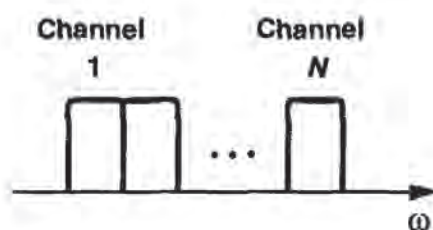


Figure 4.10 Frequency-division multiple access.

The relative simplicity of FDMA made it the principal access method in early cellular networks, called “analog FM” systems. However, in FDMA the minimum number of simultaneous users is given by the ratio of the total available frequency band and the width of each channel, often translating into insufficient capacity in crowded areas.

4.2.3 Time-Division Multiple Access

In another implementation of multiple-access networks, the same band is available to each user but at different times (time-division multiple access). Illustrated in Fig. 4.11, TDMA periodically enables each of the transceivers for a time slot (T_{sl}). The overall period consisting of all of the time slots is called a frame (T_F). In other words, every T_F seconds, each user finds access to the channel for T_{sl} seconds.

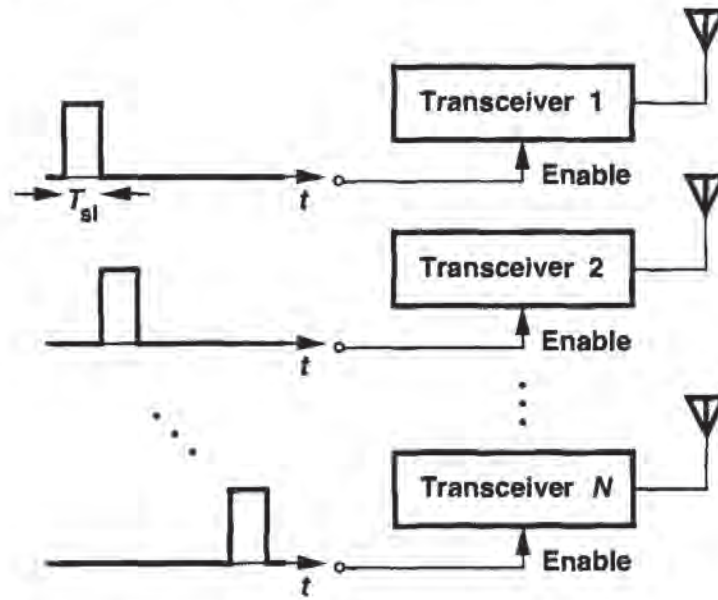


Figure 4.11 Time-division multiple access.

What happens to the data (e.g., voice) produced by all other users while only one user is allowed to transmit? To avoid loss of information, the data is stored (“buffered”) for $T_F - T_{sl}$ seconds and transmitted as a burst during one time slot (hence the term “TDMA burst”). Since buffering requires the data to be in digital form, TDMA transmitters perform A/D conversion on the analog input signal. Digitization also allows speech compression and coding.

TDMA systems have a number of advantages over their FDMA counterparts. First, as each transmitter is enabled for only one time slot in every frame, the power amplifier can be turned off during the rest of the frame, thus saving considerable power. In practice, settling issues require that the PA be turned on slightly before the actual time slot begins. Second, since digitized speech can be compressed in time by a large factor, the required communication bandwidth can be smaller and hence the overall capacity larger. Equivalently, as compressed speech can be transmitted over a shorter time slot, a higher number of users can be accommodated in each frame. Third, even with FDD, the TDMA bursts can be timed such that the receive and transmit paths in each transceiver are never enabled simultaneously.

The need for A/D conversion, digital modulation, time slot and frame synchronization, etc., makes TDMA more complex than FDMA. With the advent of VLSI DSPs, however, this drawback is no longer a determining factor.

In most actual TDMA systems, a combination of TDMA and FDMA is utilized. In other words, each of the channels depicted in Fig. 4.10 is time shared among many users.

4.2.4 Code-Division Multiple Access

Our discussion of FDMA and TDMA implies that the transmitted signals in these systems avoid interfering with each other in either the frequency domain or the time domain. In essence, the signals are orthogonal in one of these domains. A third method of multiple access allows complete overlap of signals in both frequency and time, but employs “orthogonal messages” to avoid interference. This can be understood with the aid of an analogy [3]. Suppose many conversations are going on in a crowded party. To minimize crosstalk, different groups of people can be required to speak in different pitches(!) (FDMA), or only one group can be allowed to converse at a time (TDMA). Alternatively, each group can be asked to speak in a *different language*. If the languages are orthogonal (at least in nearby groups) and the voice levels are roughly the same, then each listener can “tune in” to the proper language and receive information while all groups talk simultaneously.

Direct-Sequence CDMA In “code-division multiple access,” different languages are created by means of orthogonal digital codes. At the beginning of communication, a certain code is assigned to each transmitter/receiver pair, and each bit of the baseband data is “translated” to that code before modulation. Shown in Fig. 4.12(a) is an example, where each baseband pulse is replaced with an 8-bit code by multiplication. A method of generating orthogonal codes is based on Walsh’s recursive equation

$$W_1 = 0 \quad (4.4)$$

$$W_{2n} = \begin{bmatrix} W_n & W_n \\ W_n & \overline{W_n} \end{bmatrix}, \quad (4.5)$$

where $\overline{W_n}$ is derived from W_n by replacing all the entries with their complements. For example,

$$W_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}. \quad (4.6)$$

Fig. 4.12(b) shows examples of 8-bit Walsh codes (i.e., each row of W_8).

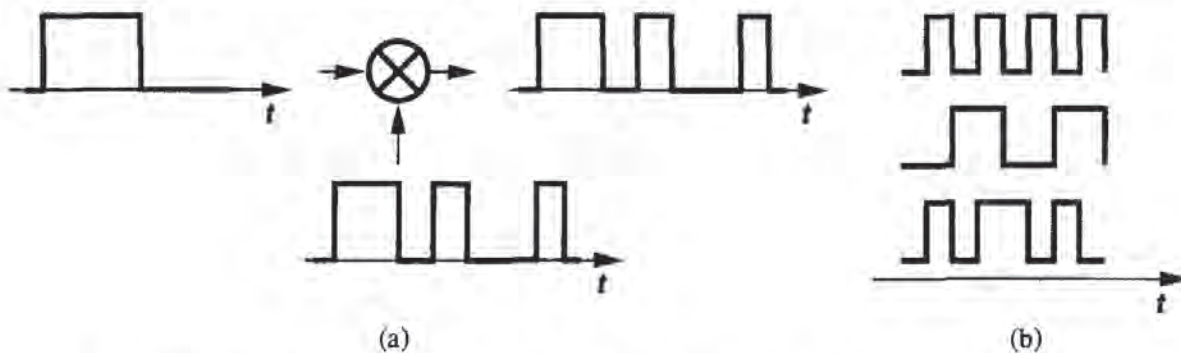


Figure 4.12 (a) Encoding operation in DS-CDMA; (b) examples of Walsh code.

In the receiver, the demodulated signal is decoded by multiplying it by the same Walsh code. In other words, the receiver *correlates* the signal with the code to recover the baseband data.

How is the received data affected when another CDMA signal is present? Suppose two CDMA signals $x_1(t)$ and $x_2(t)$ are received in the same frequency band. Writing the signals as $x_{BB1}(t) \cdot W_1(t)$ and $x_{BB2}(t) \cdot W_2(t)$, where $W_1(t)$ and $W_2(t)$ are Walsh functions, we express the output of the demodulator as $y(t) = [x_{BB1}(t) \cdot W_1(t) + x_{BB2}(t) \cdot W_2(t)] \cdot W_1(t)$. Thus, if $W_1(t)$ and $W_2(t)$ are exactly orthogonal, then $y(t) = x_{BB1}(t) \cdot W_1(t)$. In reality, however, $x_1(t)$ and $x_2(t)$ may experience different delays, leading to corruption of $y(t)$ by $x_{BB2}(t)$. Nevertheless, for long codes the result appears as random noise.

The encoding operation of Fig. 4.12(a) *increases* the bandwidth of the data spectrum by the number of pulses in the code. This may appear in contradiction to our emphasis thus far on spectral efficiency. However, since CDMA allows the widened spectra of many users to fall in the same frequency band (Fig. 4.13), this access technique has no less capacity than do FDMA and TDMA. In fact, CDMA can potentially achieve a higher capacity than the other two [4].

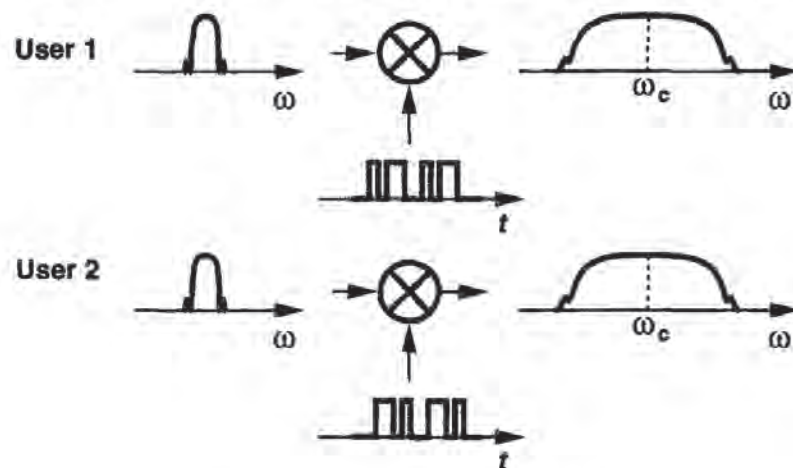


Figure 4.13 Overlapping spectra in CDMA.

CDMA is a special case of “spread spectrum” (SS) communications, whereby the baseband data of each user is spread over the entire available bandwidth. In this context, CDMA is also called “direct-sequence” SS (DS-SS) communication, and the code is called the “spreading sequence” or “pseudo-random noise.” To avoid confusion with the baseband data, each pulse in the spreading sequence is called a “chip,” and the rate of the sequence is called the “chip rate.” Thus, the spectrum is spread by the ratio of the chip rate to the baseband bit rate.

It is instructive to reexamine the above decoding operation from a spread spectrum point of view (Fig. 4.14). Upon multiplication by $W_1(t)$, the desired signal is “despread,” with its bandwidth returning to the original value. The unwanted signal, on the other hand, remains spread even after multiplication

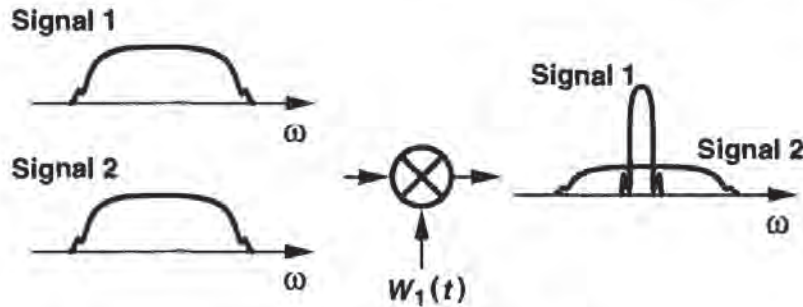


Figure 4.14 Despreading operation.

because of its low correlation with $W_1(t)$. For a large number of users, the spread spectra of unwanted signals can be viewed as white Gaussian noise.

An important feature of CDMA is its soft capacity limit [1]. While in FDMA and TDMA the maximum number of users is fixed once the channel width or the time slots are defined, in CDMA increasing the number of users only gradually (linearly) raises the noise floor [1].

A critical issue in DS-CDMA is power control. Suppose, as illustrated in Fig. 4.15, the desired signal power received at a point is much lower than that of an unwanted transmitter,¹ for example, because the latter is at a shorter distance. Even after despreading, the strong interferer greatly raises the noise floor, degrading the reception of the desired signal. For multiple users, this means that one high-power transmitter can virtually halt communications among others, a problem much less serious in FDMA and TDMA. For this reason, when many CDMA transmitters communicate with a receiver, they must adjust their output power such that the receiver senses roughly equal signal levels. To this end, the receiver monitors the signal strength corresponding to each transmitter and periodically sends a power adjustment request to each one. Since in a cellular system users communicate through the base station, rather than directly, the latter must handle the task of power control. The received signal levels are controlled to be typically within 1 dB of each other.

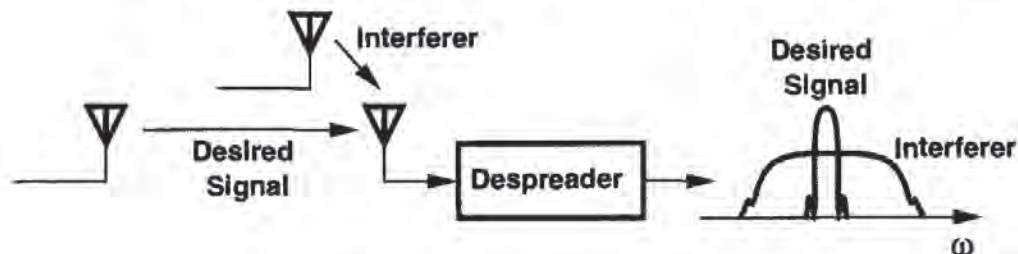


Figure 4.15 Effect of strong interferer in CDMA.

¹ This situation arises in our party analogy if two people speak much more loudly than others. Even with different languages, communication becomes difficult.

While adding complexity to the system, power control generally reduces the average power dissipation of the mobile unit. To understand this, note that without such control, the mobile must *always* transmit enough power to be able to communicate with the base station, whether path loss and fading are significant or not. Thus, even when the channel has minimum attenuation, the mobile unit produces the maximum output power. With power control, on the other hand, the mobile can transmit at low levels whenever the channel conditions improve. This also reduces the average interference seen by other users.

Frequency-Hopping CDMA Another type of CDMA that has begun to appear in RF communications is “frequency hopping” (FH). Illustrated in Fig. 4.16, this access technique can be viewed as FDMA with pseudorandom channel allocation. The carrier frequency in each transmitter is “hopped” according to a chosen code (similar to the spreading codes in DS-CDMA). Thus, even though the short-term spectrum of a transmitter may overlap with those of others, the overall trajectory of the spectrum, i.e., the PN code, distinguishes each transmitter from others. Nevertheless, occasional overlap of the spectra raises the probability of error.

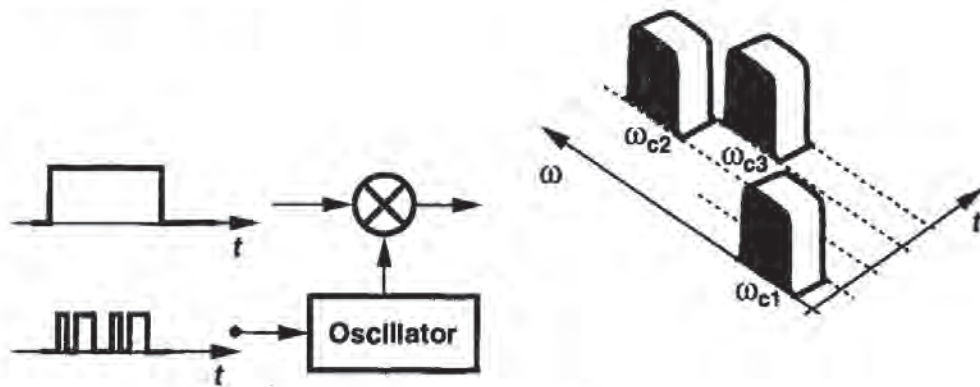


Figure 4.16 Frequency-hopping CDMA.

Due to rare overlap of spectra, frequency hopping is somewhat similar to FDMA and hence more tolerant of different received power levels than is direct-sequence CDMA. However, FH may require relatively fast settling in the control loop of the oscillator shown in Fig. 4.16, an important design issue studied in Chapter 8.

4.3 WIRELESS STANDARDS

From our study of modulation and multiple access techniques, we can see that making a phone call in a wireless system entails a great many complex operations. Furthermore, the existence of nonidealities such as noise and interference mandates a precise specification of communication parameters, for example,

SNR, BER, occupied bandwidth, and extraneous emissions. A wireless standard defines all the details and constraints that govern the design of transceivers used in a wireless system.² While providing the “boundary conditions,” such a standard does not specify how the actual transmit and receive paths must be implemented, allowing some flexibility in the choice of transceiver architectures (Chapter 5).

In this section, we study the “air interface” of a number of standards used in cellular and cordless phone systems.

4.3.1 Advanced Mobile Phone Service

Among the earliest wireless standards, Advanced Mobile Phone Service (AMPS) employs FDMA with analog FM and FDD. Partitioned into 30-kHz channels, the transmit and receive bands of the mobile phone are 824–849 MHz and 869–894 MHz, respectively [Fig. 4.17(a)]. Thus, the system can support approximately 830 users simultaneously.³ The 20-MHz separation between the two bands allows the use of relatively low loss duplexer filters [Fig. 4.17(b)].

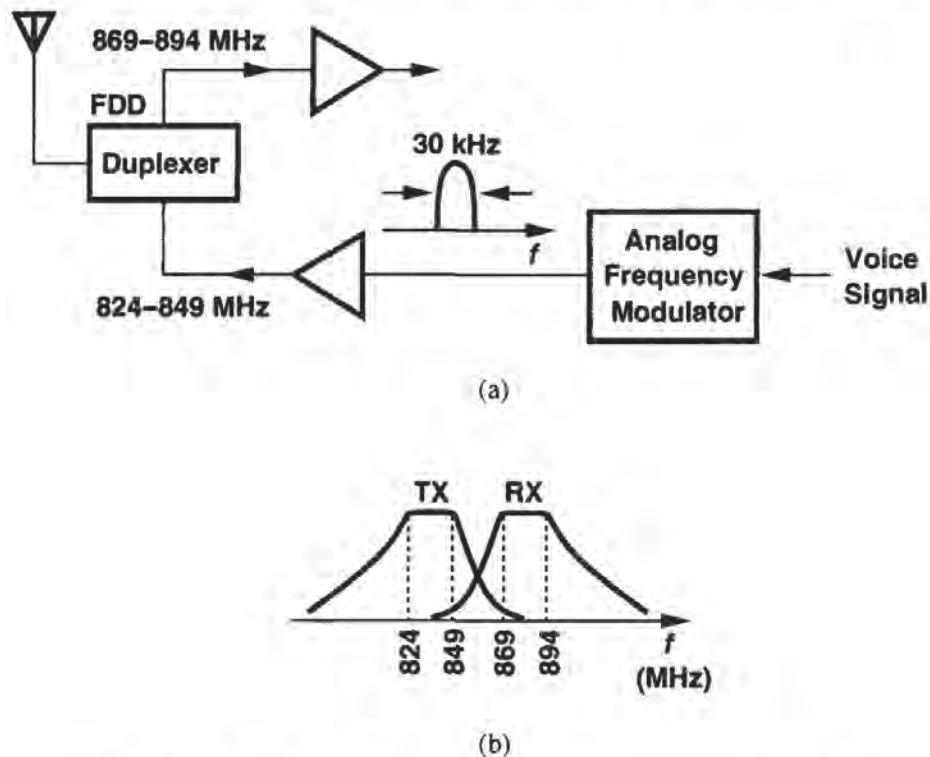


Figure 4.17 (a) AMPS air interface, (b) duplexer characteristics.

² The documentation of some standards is several thousand pages long.

³ In reality, only half of these are given to each long-distance carrier.

In addition to the voice channel, AMPS requires control and supervisory signals to initiate, maintain, and terminate a call. These signals are transmitted over the same channel and are described in [1].

4.3.2 North American Digital Standard

As the first digital cellular system in the United States, North American Digital Cellular (NADC) system employs TDMA with $\pi/4$ -DQPSK and FDD [Fig. 4.18(a)]. Designed to be compatible with AMPS, NADC uses the same transmit and receive bands with 30-kHz channel spacing while providing six times the capacity of AMPS. The bit rate in each channel is 48.6 kb/s.

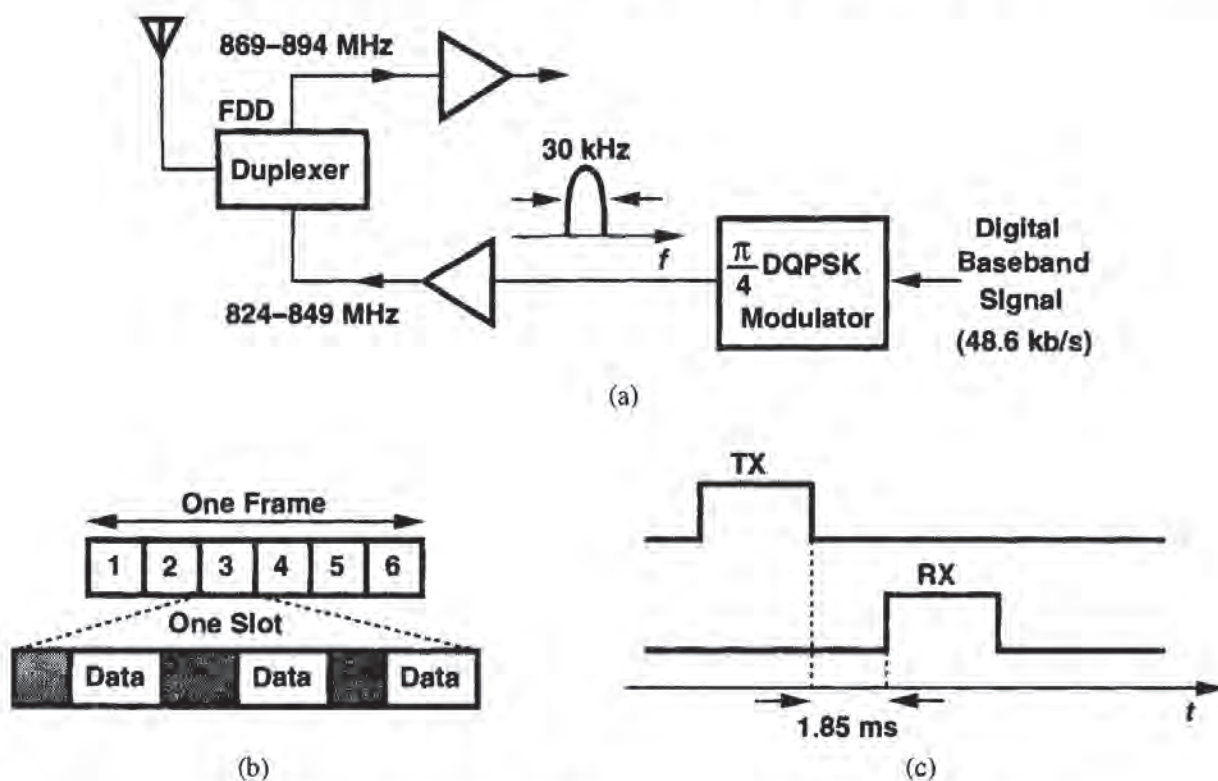


Figure 4.18 NADC (a) air interface, (b) frame structure, (c) TX and RX time slots.

The TDMA frame structure used in NADC is illustrated in Fig. 4.18(b) [1]. Each frame is 1944 bits (40 ms) long and consists of 6 time slots, and each time slot carries approximately 260 bits of data along with 64 bits of control and synchronization information. Note that, as shown in Fig. 4.18(c), the transmit and receive time slots are offset by 1.85 ms [5]. This significantly relaxes transceiver design issues such as leakage of the transmitted signal to the receive band and desensitization of the low-noise amplifier by the power amplifier.

Most first-generation digital phones in the United States actually operate with both AMPS and NADC to provide a wider coverage for users. Called dual-mode AMPS/NADC systems, these phones are designed according to

Interim Standard 54 (IS-54), developed by the Electronic Association and the Telecommunication Industry Association.

For details of control signals and interleaving in IS-54, the reader is referred to [1].

4.3.3 Global System for Mobile Communication

Originally developed as a unified wireless standard for Europe, GSM⁴ has become the most widely used cellular standard in the world. In addition to regular phone calls, GSM supports many other services such as facsimile and ISDN [1], making the mobile phone a versatile communicator.

The GSM standard is a TDMA/FDD system using GMSK modulation, with a transmit band of 890–915 MHz and a receive band of 935–960 MHz [Fig. 4.19(a)]. Accommodating eight time-multiplexed users, each channel is 200 kHz wide, and the data rate per user is 270 kb/s. The TDMA frame structure is depicted in Fig. 4.19(b). Each frame is 4.615 ms long and consists of 8 time

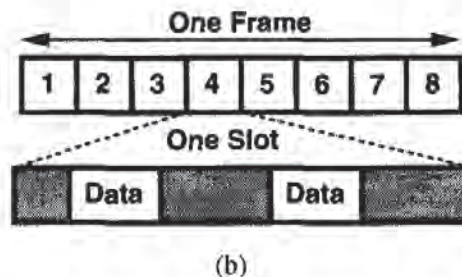
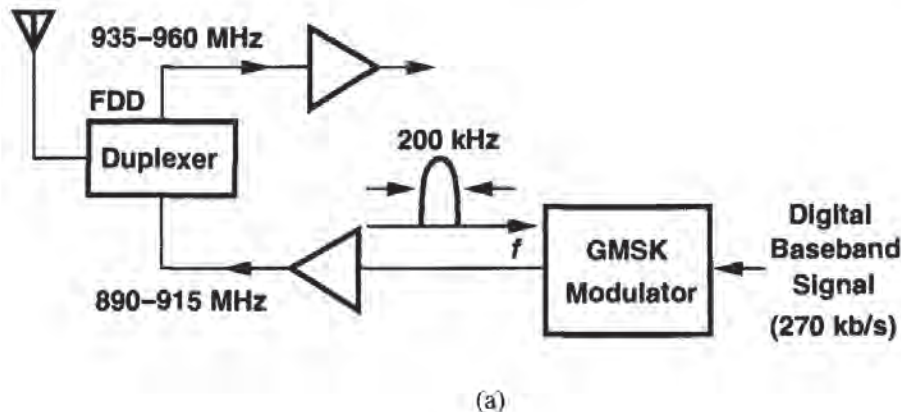


Figure 4.19 GSM (a) air interface and (b) frame structure.

⁴ GSM originally stood for Groupe Speciale Mobile.

slots each carrying 114 bits of data along with other control and training bits. Similar to NADC, the TX and RX time slots are offset by three time slots so that the transmitter and the receiver do not operate simultaneously. The total capacity of the system is given by the number of channels in the 25-MHz bandwidth and the number of users per channel, approximately 1000.

GSM employs a variety of data and control channels. The reader is referred to [1] for details.

4.3.4 Qualcomm CDMA

A wireless standard based on direct-sequence CDMA has been proposed by Qualcomm, Inc., and adopted for the North America as IS-95. Using FDD, the air interface employs the same transmit and receive bands as those in IS-54 (Fig. 4.20). In the mobile unit, the 9.6 kb/s baseband data is spread to 1.23 MHz and subsequently modulated using OQPSK. The link from the base station to the mobile unit, on the other hand, incorporates QPSK modulation. The logic is that the mobile must use a power-efficient modulation scheme (Chapter 3), whereas the base station transmits many channels simultaneously and must therefore employ a linear power amplifier regardless of the type of modulation. In both directions, IS-95 requires *coherent* detection, a task accomplished by transmitting a relatively strong “pilot tone” (e.g., unmodulated carrier) at the beginning of communication to establish phase synchronization.

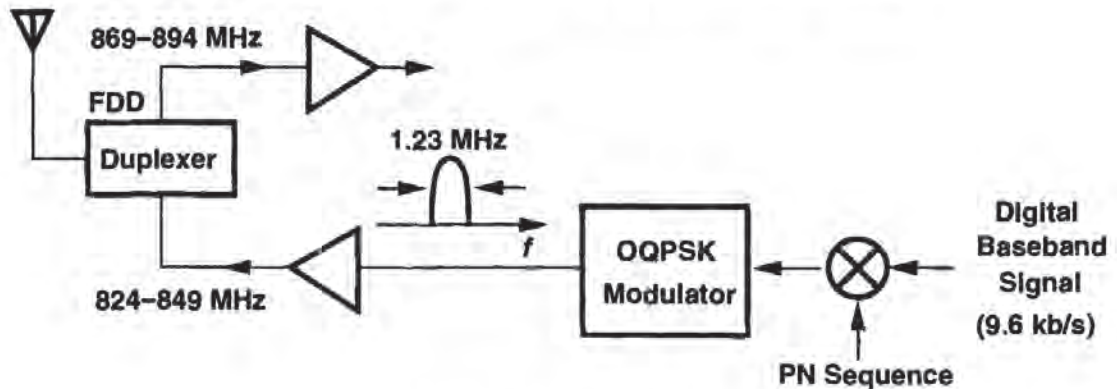


Figure 4.20 IS-95 air interface.

In contrast to the other standards studied above, IS-95 is substantially more complex, incorporating many techniques to increase the capacity while maintaining a reasonable signal quality. We briefly describe some of the features here. For more details, the reader is referred to [3, 6, 7].

Power Control As mentioned in Section 4.2.4, in CDMA the power levels received by the base station from various mobile units must differ by no more than approximately 1 dB. In IS-95, the output power of each mobile is controlled by an open-loop procedure at the beginning of communication

so as to perform a rough, but fast adjustment. Subsequently, the power is set more accurately by a closed-loop method. For open-loop control, the mobile measures the signal power it *receives* from the base station and adjusts its transmitted power so that the sum of the two (in dB) is approximately -73 dBm. If the receive and transmit paths entail roughly equal attenuation, k dB, and the power transmitted by the base station is P_{bs} , then the mobile output power P_m satisfies the following: $P_{bs} - k + P_m = -73$ dBm. Since the power received by the base station is $P_m - k$, we have $P_m - k = -73$ dBm $- P_{bs}$, a well-defined value because P_{bs} is usually fixed. The mobile output power can be varied by approximately 85 dB in a few microseconds.

Closed-loop power control is also necessary because the above assumption of equal loss in the transmit and receive paths is merely an approximation. In reality, the two paths may experience different fading because they operate in different frequency bands. To this end, the base station measures the power level received from the mobile unit and sends a feedback signal requesting power adjustment. This command is transmitted once every 1.25 ms to ensure timely adjustment in the presence of rapid fading.

Frequency and Time Diversity Recall from Section 4.1 that multipath fading is often frequency selective, causing a notch in the channel transfer function that can be several kilohertz wide. Since IS-95 spreads the spectrum to 1.23 MHz, it provides frequency diversity, exhibiting only 25% loss of the band for typical delay spreads [3].

IS-95 also employs time diversity to use multipath signals to advantage. This is accomplished by performing correlation on *delayed replicas* of the received signal (Fig. 4.21). Called a “rake receiver,” such a system combines the delayed replicas with proper weighting factors, α_j , to obtain the maximum signal-to-noise ratio at the output. That is, if the output of one correlator is corrupted, then the corresponding weighting factor is reduced and vice versa.

Rake receivers are a unique feature of CDMA. Since the chip rate is much higher than the fading bandwidth, and since the spreading codes are designed

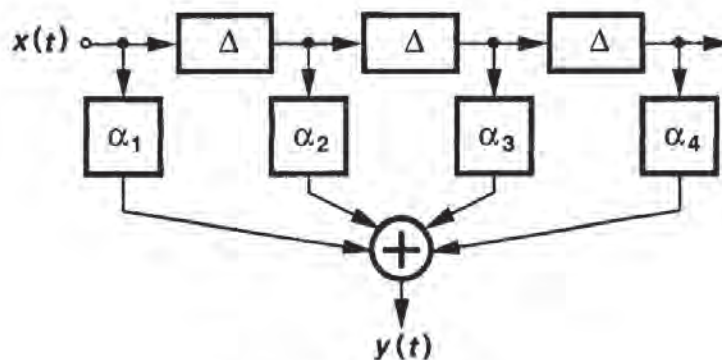


Figure 4.21 Rake receiver.

to have negligible correlation for delays greater than a chip period, multipath effects do not introduce intersymbol interference. Thus, each correlator can be synchronized to one of the multipath signals.

Variable Coding Rate The variable rate of information in human speech can be exploited to lower the average number of transmitted bits per second. In IS-95, the data rate can vary in four discrete steps: 9600, 4800, 2400, and 1200 b/s. This arrangement allows buffering slower data such that the transmission still occurs at 9600 b/s but for a proportionally shorter duration. For example, if the speech rate is 2400 b/s, then data can be accumulated for 50% of the time and transmitted only for the other 50%. This approach further reduces the average power transmitted by the mobile unit, both saving battery and lowering interference seen by other users.

Soft Handoff Recall from Section 4.1 that when the mobile unit is assigned a different base station, the call may be dropped if the channel center frequency must change (e.g., in IS-54 and GSM). In CDMA, on the other hand, all of the users in one cell communicate on the same channel. Thus, as the mobile unit moves farther from one base station and closer to another, the signal strength corresponding to *both* stations can be monitored by means of a rake receiver. When it is ascertained that the nearer base station has a sufficiently strong signal, the hand-off is performed. Called “soft handoff,” this method can be viewed as a “make-before-break” operation. The result is lower probability of dropping calls during handoff.

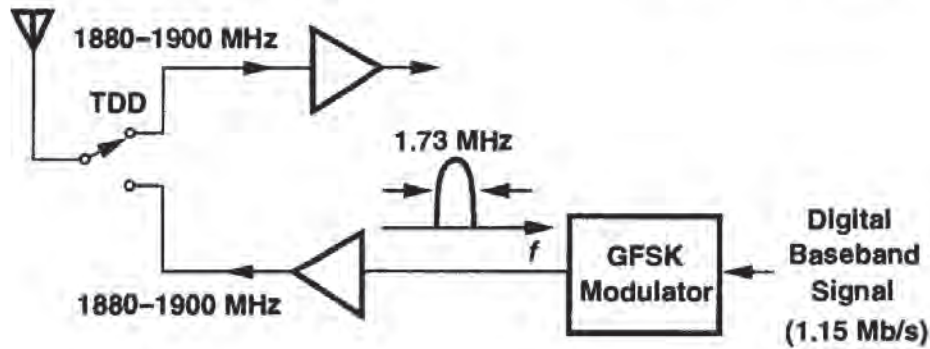
4.3.5 Digital European Cordless Telephone

The Digital European Cordless Telephone (DECT) standard was originally adopted as a cordless phone framework in Europe. The flexibility of this standard, however, has made it attractive for a wide range of applications. Designed to allow connection to other systems such as GSM, DECT provides mobility to local area network users, for example, in buildings.

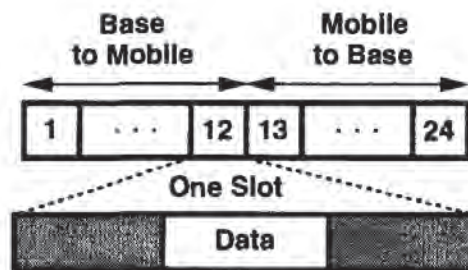
Using TDMA/FDMA, DECT operates with TDD in the frequency band 1880 MHz to 1900 MHz [Fig. 4.22(a)]. Time-shared among eight users, each channel is 1.73 MHz wide, yielding a total capacity of 120. The modulation scheme is Gaussian frequency shift keying (GFSK), a type of GMSK.

Shown in Fig. 4.22(b) is the DECT TDMA frame structure. It consists of 24 time slots (12 for transmit and 12 for receive) with a total duration of 10 ms. Each time slot contains 32 preamble bits, 388 data bits, and 60 guard bits.

In contrast to cellular standards, DECT aims at low cost and low power in the design of the portable phone rather than spectral efficiency. For example, the wide channel spacing simplifies the design of frequency synthesizers (Chapter 8).



(a)



(b)

Figure 4.22 DECT (a) air interface and (b) frame structure.

REFERENCES

- [1] T. S. Rappaport, *Wireless Communications, Principles and Practice*, Englewood Cliffs, NJ: Prentice Hall, 1996.
- [2] L. W. Couch, *Digital and Analog Communication Systems*, 4th ed., New York: Macmillan, 1993.
- [3] D. P. Whipple, "North American Cellular CDMA," *Hewlett-Packard Journal*, pp. 90-97, December 1993.
- [4] A. Salmasi and K. S. Gilhousen, "On the System Design Aspects of Code Division Multiple Access (CDMA) Applied to Digital Cellular and Personal Communications Networks," *Proc. IEEE Veh. Tech. Conf.*, pp. 57-62, May 1991.
- [5] T. W. Baker and P. E. Bronner, "Create a Low-Cost Cellular Telephone," *Wireless Systems Design*, pp. 23-33, January 1996.
- [6] R. Kerr et al, "The CDMA Digital Cellular System, An ASIC Overview," *Proceedings of IEEE CICC*, pp. 10.1.1-10.1.7, May 1992.
- [7] J. Hinderling et al, "CDMA Mobile Station Modem ASIC," *Proceedings of IEEE CICC*, pp. 10.2.1-10.2.5, May 1992.

TRANSCEIVER ARCHITECTURES

Our study of multiple access techniques and standards in Chapter 4 treated wireless systems at the mobile station level. In this chapter, we move down to the transceiver architecture level. Complexity, cost, power dissipation, and the number of external components have been the primary criteria in selecting transceiver architectures. As IC technologies evolve, however, the relative importance of each of these criteria changes, allowing approaches that once seemed impractical to return as plausible solutions.

In this chapter, we describe heterodyne, homodyne, image-reject, digital-IF, and subsampling receivers and direct-conversion and two-step transmitters. While, at present, only a few of these architectures are used in actual products, a detailed treatment of their design issues and trade-offs helps determine their potential for emerging applications.

To provide a view of the state of the art, we conclude this chapter with a study of five specific RF products currently used in wireless systems.

5.1 GENERAL CONSIDERATIONS

The wireless communications environment, especially in urban areas, is often called “hostile” because it imposes severe constraints upon transceiver design. Perhaps the most important constraint is the limited spectrum allocated to each user (e.g., 30 kHz in IS-54 and 200 kHz in GSM). From Shannon’s theorem, this translates to a limited rate of information, mandating the use of sophisticated techniques such as coding, compression, and bandwidth-efficient modulation, even for voice signals.

The narrow bandwidth available to each user also impacts the design of the RF section. As depicted in Fig. 5.1, the transmitter must employ narrowband modulation, amplification, and filtering to avoid leakage to adjacent channels, and the receiver must be able to process the desired channel while sufficiently rejecting strong neighboring interferers.

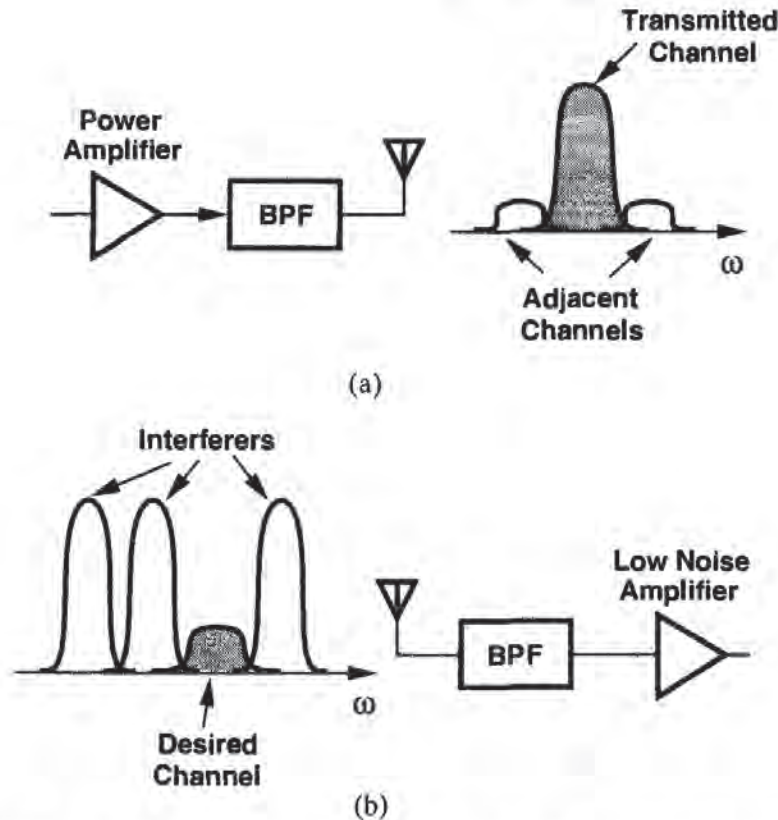


Figure 5.1 (a) Transmitter and (b) receiver front ends of a wireless transceiver.

To gain a better feeling about the difficulty of interference rejection, suppose a 900-MHz receiver incorporates a bandpass filter to select a 30-kHz channel while rejecting interfering channels 60 kHz away (Fig. 5.2). If a simple second-order LC filter is to provide 60 dB of attenuation at 45 kHz from the center, then its equivalent Q is on the order of 10^7 , a value difficult to achieve even in such filters as surface acoustic wave (SAW) devices. It is important to note that typical filters exhibit a trade-off between the loss and the Q . Now recall from Chapter 2 that a lossy circuit “magnifies” the noise figure of the succeeding stage by the attenuation factor. For example, if in Fig. 5.1(b) the

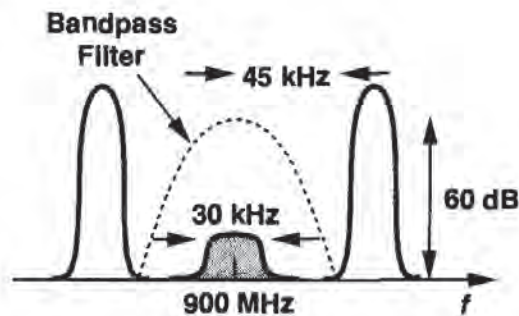


Figure 5.2 Rejection required of a hypothetical front-end bandpass filter.

bandpass filter has a loss of 2 dB and the LNA a noise figure of 2 dB, the overall noise figure rises to 4 dB. Thus, the choice of the BPF is governed by both the out-of-band rejection and the in-band loss, with the latter being the more critical parameter.

It is important to distinguish between the *band* and the *channel*: the former includes the entire spectrum in which the users of a particular standard are allowed to communicate (e.g., the GSM receive band spans 935 MHz to 960 MHz), whereas the latter refers to the signal bandwidth of only one user in the system (e.g., 200 kHz in GSM). We also apply the terms “band selection” and “channel selection” to the operations that reject out-of-band interferers and out-of-channel (usually in-band) interferers, respectively. The above calculation of the Q indicates that a practical front-end BPF can only select the band of interest, postponing the task of channel selection to some other point in the receiver (Fig. 5.3).

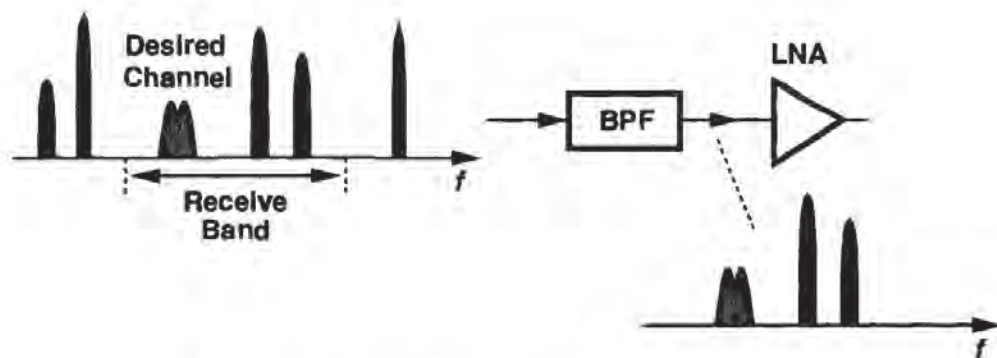


Figure 5.3 Band selection at the front end of a receiver.

We should also note that the transfer function of front-end filters has a finite transition bandwidth. Shown in Fig. 5.4 is an example where the out-of-band rejection at 20 MHz offset with respect to the receive band is approximately equal to 30 dB. In other words, an interferer appearing at this frequency is attenuated by only 30 dB, a critical issue in the design of both the receive path and the frequency synthesizer (Chapter 8).

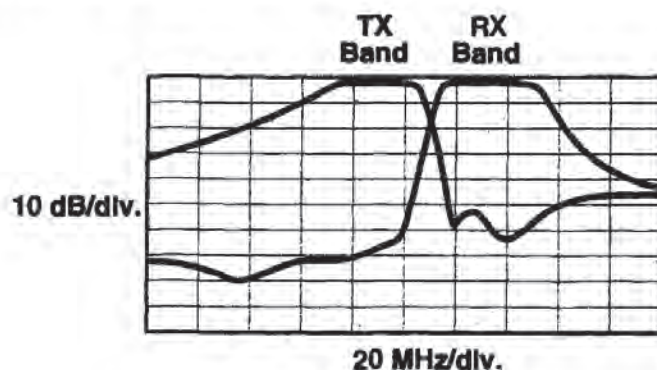


Figure 5.4 Typical duplexer characteristic.

Since large in-band interferers accompany the received signal even after the front-end BPF, the nonlinearity of the following stages, particularly that of the low-noise amplifier and the mixer, becomes important. As explained in Chapter 2 and illustrated in Fig. 5.5, odd-order nonlinearities yield intermodulation products that fall in the desired channel. As third-order distortion is usually dominant, the IP_3 of each stage must be sufficiently high to avoid corruption of the signal by the intermodulation products. Although distorting the *amplitude*, this effect is important even if the signal carries information only in its phase or frequency, because the zero-crossing points of the desired signal are corrupted by the intermodulation product.

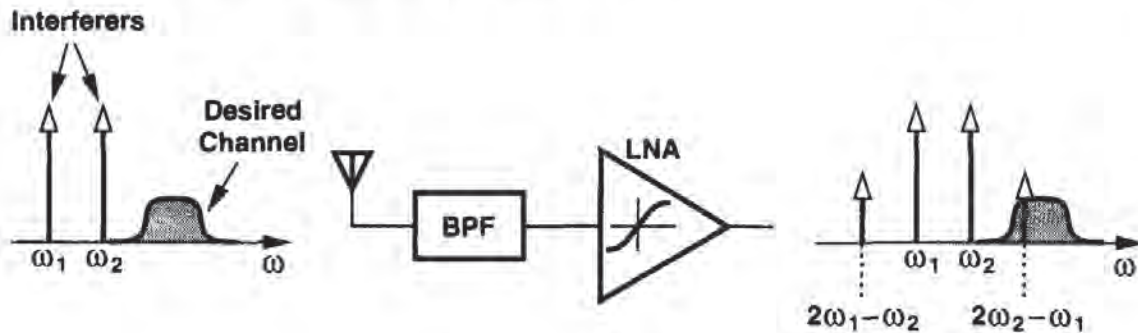


Figure 5.5 Effect of nonlinearity in the front end.

The loss of the front-end BPF also impacts the transmit path. If the power amplifier generates 1 W, then an attenuation of 2 dB in the filter translates to 370 mW loss of power, more than the typical power consumed by the entire receive path! The BPF must therefore exhibit minimal in-band loss while adequately suppressing the harmonics and out-of-band spurious components of the transmitted signal.

The above observations also indicate the importance of controlled spectral regrowth through proper choice of the modulation scheme and the power amplifier (Chapter 3). The out-of-channel intermodulation products created by the PA cannot be suppressed by the BPF and must be acceptably small by design.

Another important concern in the design of transceivers is the dynamic range of the signals. With multipath fading and path loss, the required dynamic range for the received signal is typically greater than 100 dB. As the minimum detectable signal is in the microvolt range, not only the input noise of the receiver but also cross-talk become critical. An interesting example arises in FDD systems when the finite attenuation of the transmitted signal in the receive band is considered. As illustrated in Fig. 5.6, if the TX power amplifier delivers 1 W to a 50- Ω antenna, the peak-to-peak voltage swing at the antenna is equal to 20 V. Thus, from the duplexer characteristic of Fig. 5.4, we note that the leakage to the receive path is on the order of 30 mV_{pp} (≈ -26 dBm). Since the LNA 1-dB compression point is in the vicinity of -25 dBm, the leakage

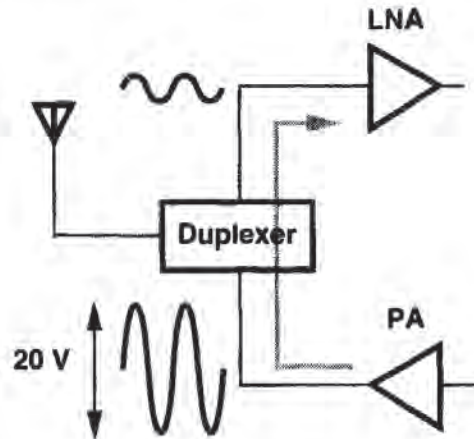


Figure 5.6 Desensitization of LNA by PA output leakage.

signal may significantly desensitize the LNA. As explained in Chapter 4, NADC and GSM systems avoid this issue by offsetting the transmit and receive time slots, but analog FDD standards such as AMPS require high isolation.

At the other extreme of the dynamic range, the receiver may experience large signals, for example, amplitudes as high as several hundred millivolts, if it is close to a transmitter. While the signal amplitude is not critical per se in PM and FM systems, the receive path must still process the signal correctly. This issue often leads to the use of automatic gain control (AGC).

The last general issue relates to power amplifiers. In recent generations of RF transceivers, the PA is periodically turned on and off to save power. However, the large current drawn by the PA (a peak value of several amperes) introduces tremendous noise in the supplies and, with typical battery output impedances, may change the battery voltage by several hundred millivolts. For this reason, noise immunity and supply rejection of all of the building blocks become important.

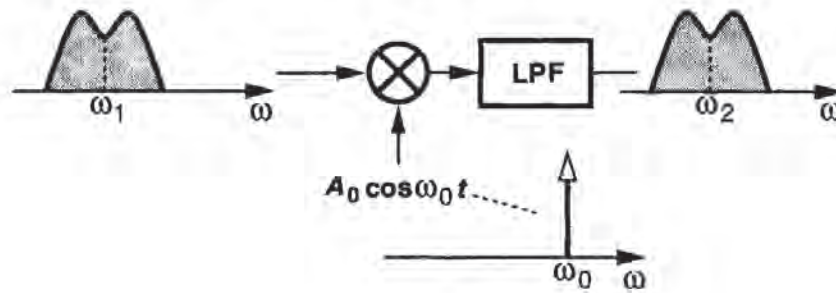
5.2 RECEIVER ARCHITECTURES

5.2.1 Heterodyne Receivers

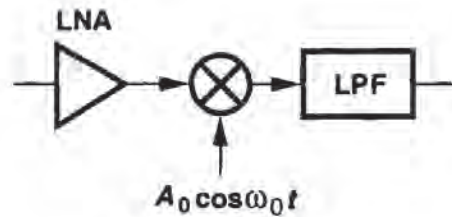
As mentioned above, filtering a narrow channel that is centered at high frequencies and is accompanied by large interferers demands prohibitively high Q 's. In heterodyne¹ architectures, the signal band is translated to much lower frequencies so as to relax the Q required of the channel-select filter. Illustrated in Fig. 5.7(a), the translation is carried out by means of a mixer, which in this chapter is viewed as a simple analog multiplier. To bring the center frequency from ω_1 to ω_2 , the signal is first mixed with a sinusoid $A_0 \cos \omega_0 t$,

¹ In this book, we do not distinguish between heterodyne and superheterodyne architectures. The term *heterodyne* derives from *hetero* (different) and *dyne* (to mix).

where $\omega_0 = \omega_1 - \omega_2$, thereby yielding a band around ω_2 and another around $2\omega_1 - \omega_2$. A low-pass filter then removes the latter. This operation is called “downconversion mixing” or simply “downconversion.” Because of its typically high noise, the downconversion mixer is preceded by a low-noise amplifier [Fig. 5.7(b)]. The sinusoid is generated by a local oscillator, and its frequency will be hereafter denoted by ω_{LO} ($= \omega_0$ in Fig. 5.7).



(a)



(b)

Figure 5.7 (a) Simple heterodyne downconversion, (b) inclusion of an LNA to lower the noise figure.

Called the “intermediate frequency” (IF), the center of the downconverted band (ω_2) is a critical parameter, bearing trade-offs with many other aspects of the performance. This will be explained below.

Problem of Image How are the LO frequency and the IF chosen? The principal consideration here is the “image frequency.” To understand the issue, note that a simple analog multiplier does not preserve the *polarity* of the difference between its two input frequencies, i.e., for $x_1(t) = A_1 \cos \omega_1 t$ and $x_2(t) = A_2 \cos \omega_2 t$, the low-pass filtered product of $x_1(t)$ and $x_2(t)$ is of the form $\cos(\omega_1 - \omega_2)t$, no different from $\cos(\omega_2 - \omega_1)t$. Thus, in a heterodyne architecture, the bands symmetrically located above and below the LO frequency are downconverted to the same center frequency (Fig. 5.8). If the received band of interest is centered around $\omega_1 (= \omega_{LO} - \omega_{IF})$, then the image is around $2\omega_{LO} - \omega_1 (= \omega_{LO} + \omega_{IF})$ and vice versa.

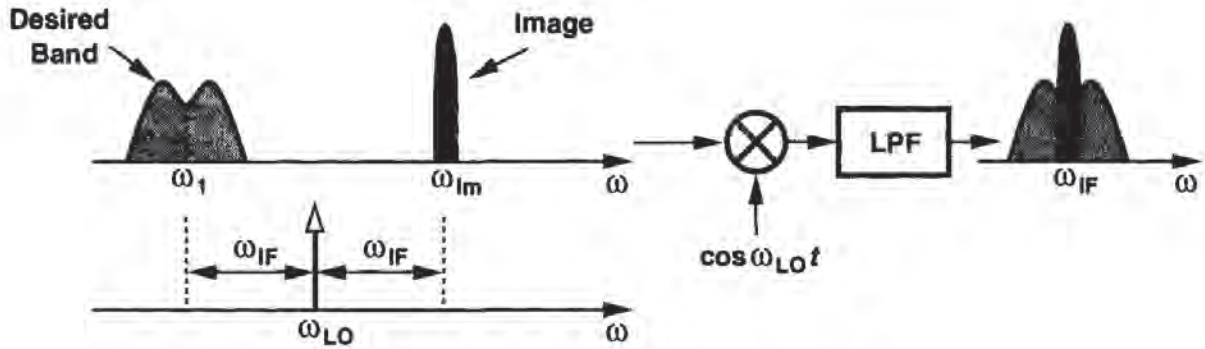


Figure 5.8 Problem of image in heterodyne reception.

The problem of image is a serious one. While each wireless standard imposes constraints upon the signal emissions by its own users, it may have no control over the signals in other bands. The image power can therefore be much higher than that of the desired signal, requiring proper “image rejection.”

The most common approach to suppressing the image is through the use of an “image-reject filter,” placed before the mixer. As depicted in Fig. 5.9, the filter is designed to have a relatively small loss in the desired band and a large attenuation in the image band, two requirements that can be simultaneously met if $2\omega_{IF}$ is sufficiently large.

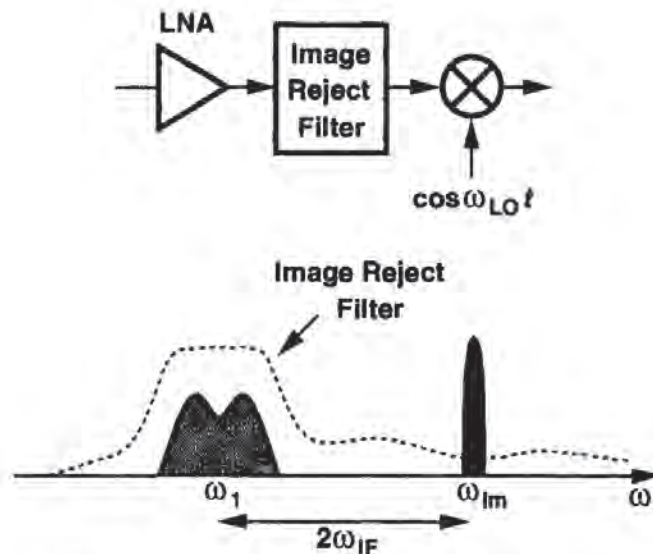


Figure 5.9 Image rejection by means of a filter.

How large can $2\omega_{IF}$ be? Recall that the premise in a heterodyne architecture is to translate the center frequency to a sufficiently low value so that channel selection by means of practical filters becomes feasible. However, as $2\omega_{IF}$ increases, so does the center of the downconverted band, requiring a higher Q in the IF filter. Shown in Fig. 5.10 are two cases corresponding to high and low values of IF so as to illustrate the trade-off. A high IF leads to

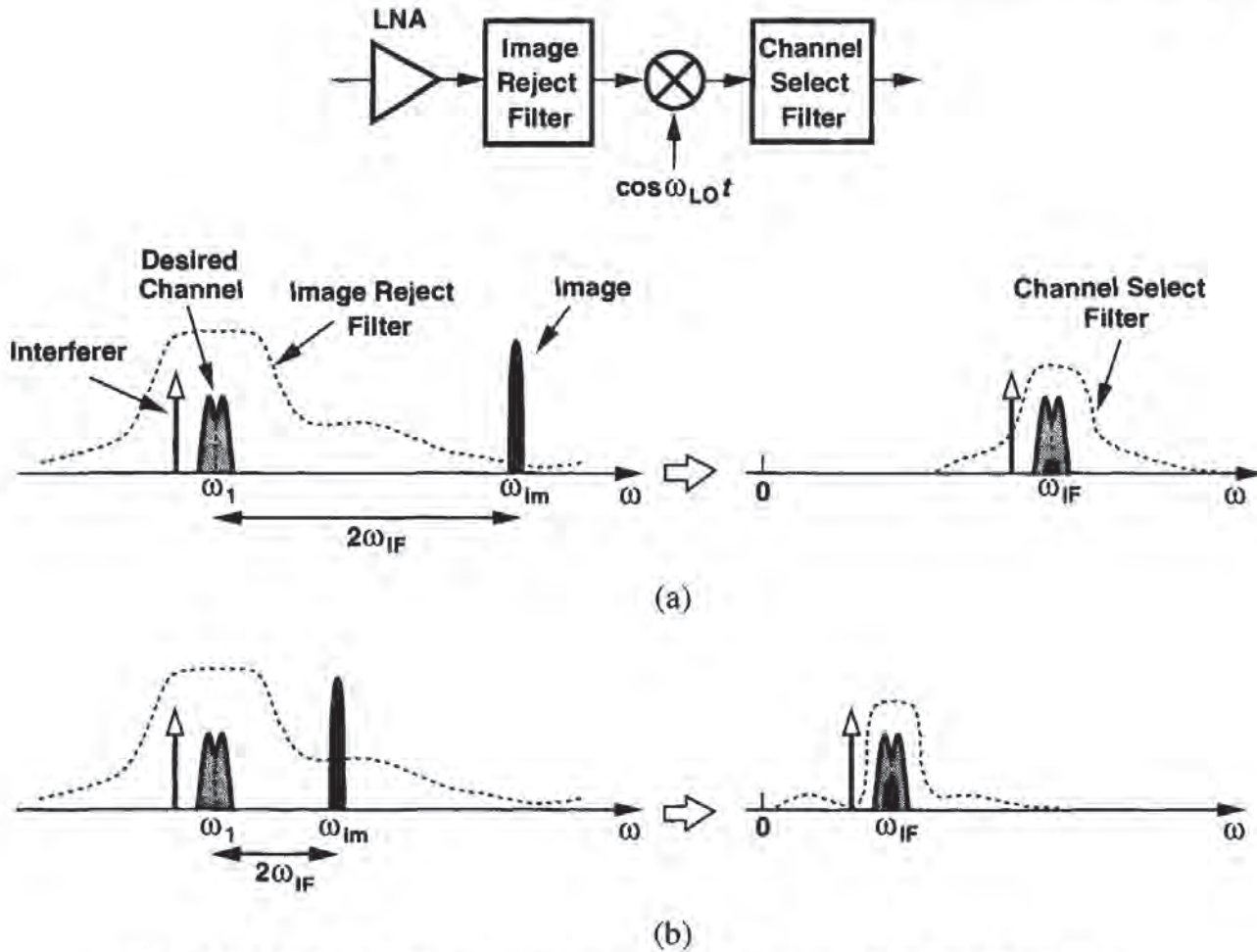


Figure 5.10 Rejection of image versus suppression of interferers for (a) high IF and (b) low IF.

substantial rejection of the image whereas a low IF allows great suppression of nearby interferers. The choice of IF therefore depends on trade-offs among three parameters: the amount of image noise, the spacing between the desired band and the image, and the *loss* of the image-reject filter. To minimize the image, we can either increase the IF or tolerate greater loss in the filter while increasing its Q . Since the gain of the LNA is typically less than 15 dB, the filter loss should not exceed a few dB, reducing the trade-off to one between the image noise and the value of IF.

From the above discussion, we infer that the heterodyne architecture of Fig. 5.9 exhibits a trade-off between image rejection and channel selection. Since the image degrades the sensitivity of the receiver, we say the choice of the IF entails a trade-off between sensitivity and selectivity.

Two other factors influence the choice of the IF: the availability and the physical size of filters for different frequencies. Employing SAW or crystal devices, these filters have been used at certain frequencies, e.g., 10.7 MHz (in

FM radios), 71 MHz, etc., and hence can be less costly than if they must be designed for other frequencies. Furthermore, to reduce the form factor of portable systems, smaller filters and hence higher IFs are more attractive.

An important drawback of the heterodyne architecture is that the image-reject filter is usually realized as a passive, external component. This furthermore requires that the preceding stage—the LNA—drive the $50\text{-}\Omega$ input impedance of the filter, inevitably leading to more severe trade-offs between the gain, noise figure, stability, and power dissipation in the amplifier (Chapter 6). In FDD systems, the duplexer may also serve to reject the image if the IF is high, thereby allowing direct connection of the LNA to the mixer.

The local oscillator frequency, ω_{LO} , can be either higher or lower than the center of the desired band. Called “high-side injection” and “low-side injection,”² respectively, the two cases entail different issues. On the one hand, it is desirable to use the latter so as to minimize the LO frequency and hence facilitate the design of the oscillator (Chapter 7). On the other hand, if the image bands below and above the desired band exhibit different amounts of noise (for example, one is used for low signal level communications and the other for a high-power wireless standard), then ω_{LO} must be chosen so as to avoid the noisy image band.

Problem of Half IF An interesting effect in heterodyne receivers is the “half IF” problem [3]. Suppose, as shown in Fig. 5.11, in addition to the desired band around ω_{in} , an interferer at $(\omega_{in} + \omega_{LO})/2$, i.e., half of the IF from the desired band toward the LO frequency, is also received. If in the downconversion path, the interferer experiences second-order distortion and the LO contains a significant second harmonic as well, then the IF output exhibits a component at $|(\omega_{in} + \omega_{LO}) - 2\omega_{LO}| = \omega_{IF}$. Another possibility is that the interferer is translated to $(\omega_{in} - \omega_{LO})/2 = \omega_{IF}/2$ and subsequently undergoes second-order distortion in the baseband, causing its second harmonic to fall into the downconverted band of interest.

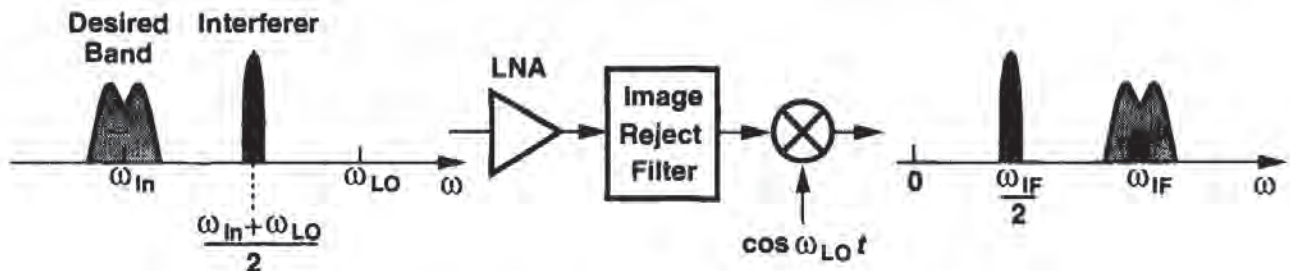


Figure 5.11 Problem of half IF in heterodyne reception.

² These have also been called “supradyn” and “infradyne,” respectively.

In order to suppress the half-IF phenomenon, second-order distortion in the RF and IF paths must be minimized, and a 50% LO duty cycle must be maintained. It may also be necessary that the image-reject filter achieve sufficient attenuation at $(\omega_{in} + \omega_{LO})/2$.

Dual-IF Topology The trade-off between sensitivity and selectivity in the simple heterodyne architecture of Fig. 5.7 often proves quite severe: if the IF is high, the image can be suppressed but complete channel selection is difficult, and vice versa. To resolve this issue, the concept of heterodyning can be extended to multiple downconversions, each followed by filtering and amplification. Illustrated in Fig. 5.12, this technique performs *partial* channel selection at progressively lower center frequencies, thereby relaxing the Q required of

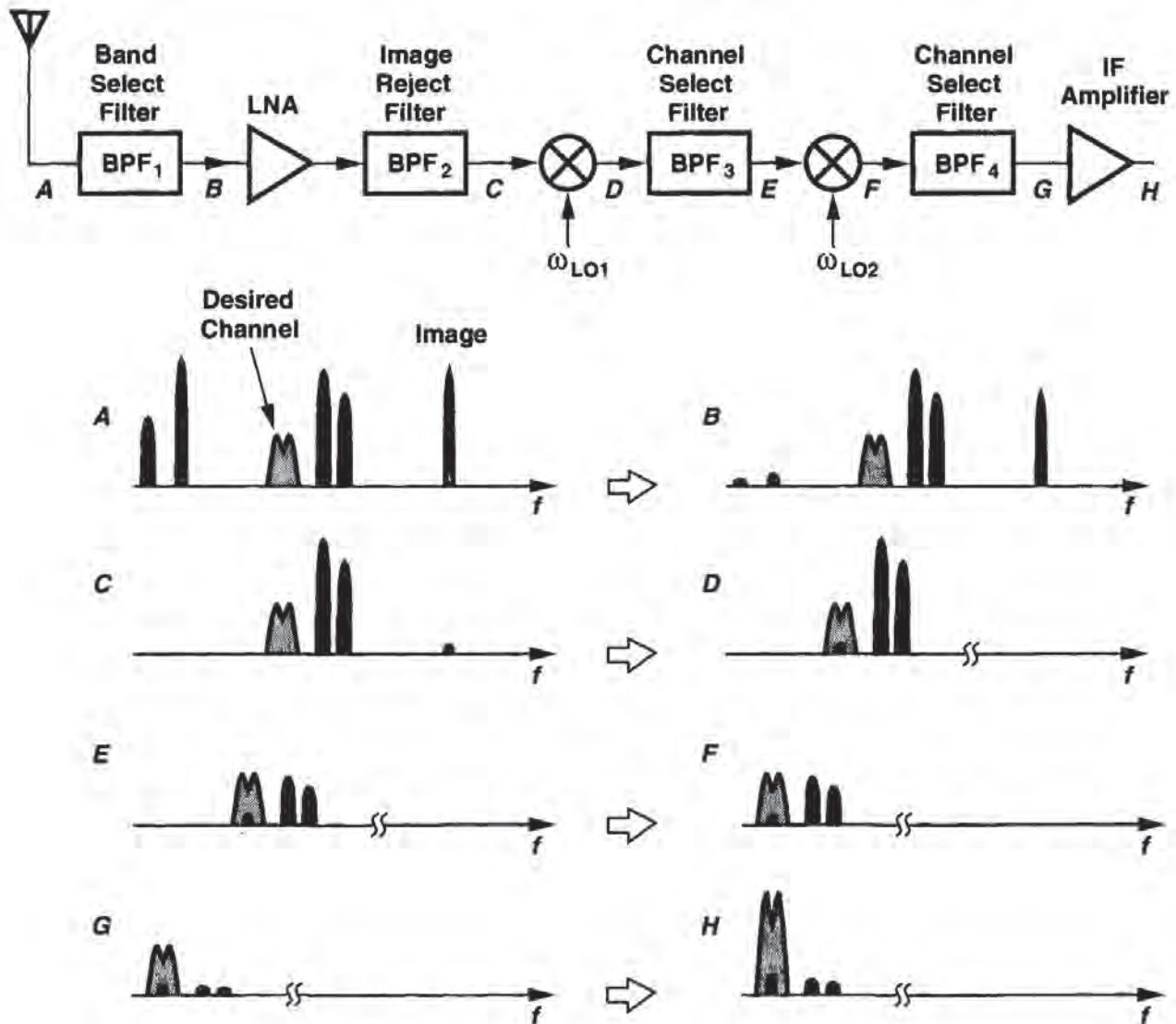


Figure 5.12 Dual-IF heterodyne receiver.

each filter. Most of today's RF receivers employ two stages of downconversion, hence the name "dual-IF." Note that the second downconversion entails the problem of image as well. For narrow-channel standards such as IS-54, the second IF is often equal to 455 kHz whereas for wide-channel applications such as DECT, it may be several megahertz. Nevertheless, these numbers vary greatly in present systems.

Recall from Chapter 2 that in a cascade of gain stages the noise figure is most critical in the front end and the linearity in the back end. Thus, an optimum design scales both the noise figure and the IP_3 of each stage according to the total gain preceding that stage. Now suppose the receiver of Fig. 5.12 exhibits a total gain of, say, 40 dB from A to G . If the two IF filters provided *no* channel selection, the IP_3 of the IF amplifier would need to be more than 40 dB higher than that of the LNA, e.g., in the vicinity of +30 dBm. However, it is difficult to achieve such high linearity with reasonable noise, power dissipation, and gain, especially if the circuit must operate from a low supply voltage. In practice, since each IF filter suppresses adjacent channel interferers to some extent, the linearity required of the stages following that filter is relaxed proportionally. This is sometimes loosely stated as "every dB of gain requires one dB of prefiltering."

Shown in Fig. 5.12 are the spectra at different points in the dual-IF receiver. The front-end filter selects the band while providing some image rejection as well. After amplification and image-reject filtering, the spectrum at point C is obtained. A sufficiently linear mixer then translates the desired channel and the adjacent interferers to the first IF. Partial channel selection in BPF_3 permits the use of a second mixer with reasonable linearity. Next, the spectrum is translated to the second IF, and BPF_4 suppresses the interferers to acceptably low levels.

What happens at the second IF? In analog FM systems, the demodulation is performed at this frequency, thus reproducing the analog baseband signal. In digital modulation systems, the second downconversion typically generates both in-phase (I) and quadrature (Q) components of the signal while translating the spectrum to zero frequency (Fig. 5.13). (This is sometimes called a "single-IF" topology because it does not include a second *intermediate* frequency.)

Since the received signal suffers from multipath fading and ISI, it is converted to digital form after I/Q detection for further processing. Depending on the interference rejection of the filters and the gain range provided by the AGC circuit, the resolution of A/D converters required here varies between approximately 4 and 10 bits.

Our analysis of heterodyne architectures indicates that the choice of the NF , IP_3 , and gain of each stage in the receive path depends on those of the preceding and following stages, thereby demanding considerable iteration at the architecture and circuit levels to arrive at an acceptable distribution of gain in the receiver building blocks. Moreover, as described in Chapter 6, each mixer generates many spurious components whose frequencies are related to those

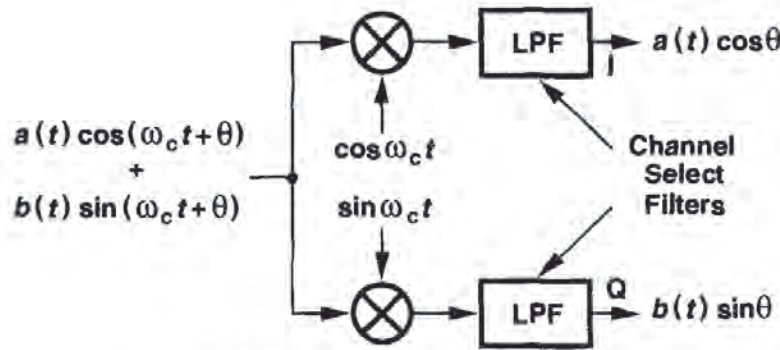


Figure 5.13 Quadrature downconversion.

of the RF and IF signals and the oscillators. Some of these components may fall in the desired channel, degrading the signal quality. Thus, the “frequency planning” of the receiver plays a key role in the overall performance that can be achieved.

The selectivity and sensitivity of the heterodyne architecture have made it the dominant choice in RF systems for many decades. Despite the complexity and the need for a large number of external components, heterodyning is still viewed as the most reliable reception technique.

5.2.2 Homodyne Receivers

In our study of heterodyne receivers, the reader may have wondered why the RF spectrum is not simply translated to the baseband in the first downconversion. Called “homodyne,” “direct-conversion,” or “zero-IF” architecture,³ this type of receiver entails vastly different issues from heterodyne topologies.

Shown in Fig. 5.14 is a simple homodyne receiver, where the LO frequency is equal to the input carrier frequency. Note that channel selection requires only a low-pass filter with relatively sharp cutoff characteristics. The circuit of Fig. 5.14(a) operates properly only with double-sideband AM signals because it overlaps positive and negative parts of the input spectrum. For frequency- and phase-modulated signals, the downconversion must provide quadrature outputs [Fig. 5.14(b)] so as to avoid loss of information. This is because the two sides of FM or QPSK spectra carry different information and must be separated into quadrature phases in translation to zero frequency.

The simplicity of the homodyne architecture offers two important advantages over a heterodyne counterpart. First, the problem of image is circumvented because $\omega_{IF} = 0$. As a result, no image filter is required, and the LNA need not drive a 50- Ω load. Second, the IF SAW filter and subsequent down-conversion stages are replaced with low-pass filters and baseband amplifiers that are amenable to monolithic integration.

³ The term *homodyne*, stemming from *homo* (same) and *dyne* (to mix), has been historically used for only coherent reception. We do not make this distinction here.

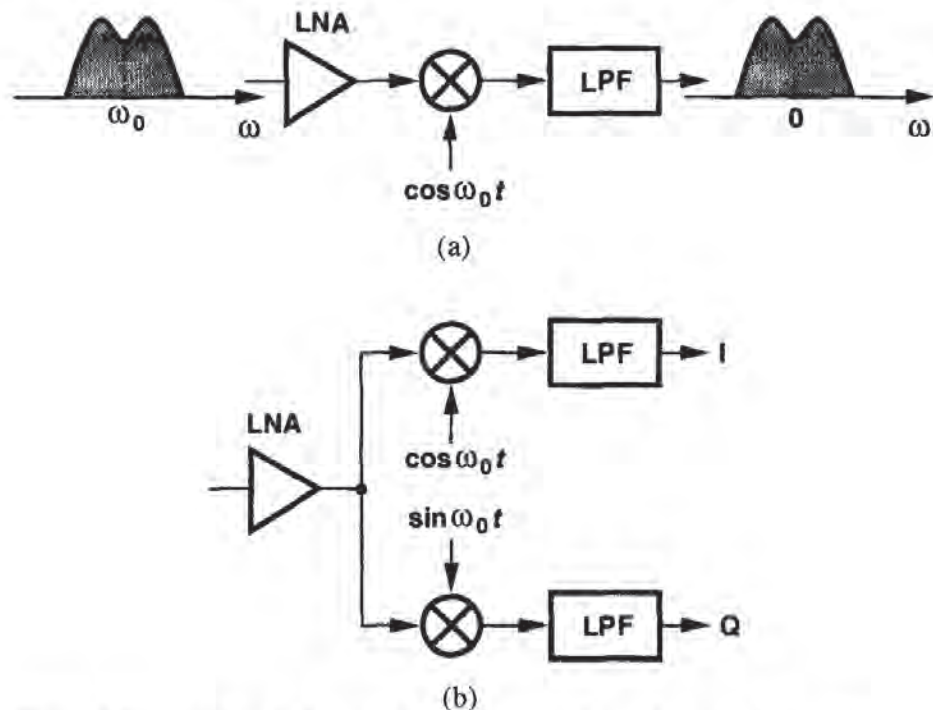


Figure 5.14 (a) Simple homodyne receiver, (b) homodyne receiver with quadrature downconversion.

If the homodyne architecture is so simple, why has it not become popular in RF systems? Direct translation of the spectrum to zero frequency entails a number of issues that do not exist or are not as serious in a heterodyne receiver.

Channel Selection Rejection of out-of-channel interferers by an active low-pass filter is more difficult than by a passive filter, fundamentally because active filters exhibit much more severe noise-linearity-power trade-offs than do their passive counterparts. The baseband processing in each branch of Fig. 5.14(b) can assume one of the three permutations shown in Fig. 5.15 [6].

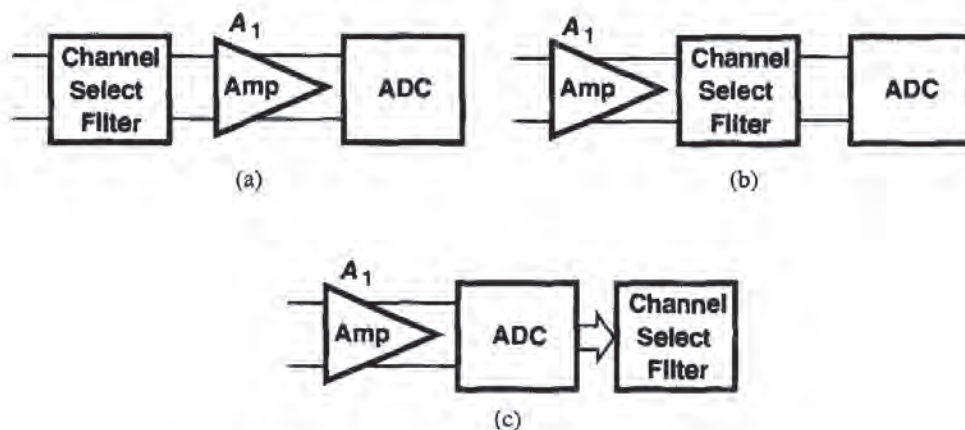


Figure 5.15 Three permutations of baseband functions.

In Figure 5.15(a), a low-pass filter suppresses out-of-channel interferers, allowing A_1 to be a nonlinear, high-gain amplifier and the ADC to have a moderate dynamic range (roughly 4 to 8 bits depending on the gain control in the RF domain and the type of modulation.) However, the low-pass filter imposes tight noise-linearity trade-offs.

The second permutation, shown in Figure 5.15(b), relaxes the LPF noise requirements while demanding a higher performance in the amplifier. A linearized one-stage differential amplifier may be employed here to provide some gain before channel filtering. Furthermore, another amplifier may be interposed between the LPF and the ADC to overcome the noise of the latter.

The third permutation, Figure 5.15(c), suggests the possibility of channel filtering in the digital domain. In this case, the ADC must both achieve a high linearity so as to digitize the signal with minimal intermodulation of interferers and exhibit a thermal and quantization noise floor well below the signal level, which may be in the range of a few hundred microvolts.

DC Offsets Since in a homodyne topology the downconverted band extends to zero frequency, extraneous offset voltages can corrupt the signal and, more importantly, saturate the following stages. To understand the origin and impact of offsets, consider the receiver shown in Fig. 5.16, where the LPF is followed by an amplifier and an A/D converter. Let us make two observations.

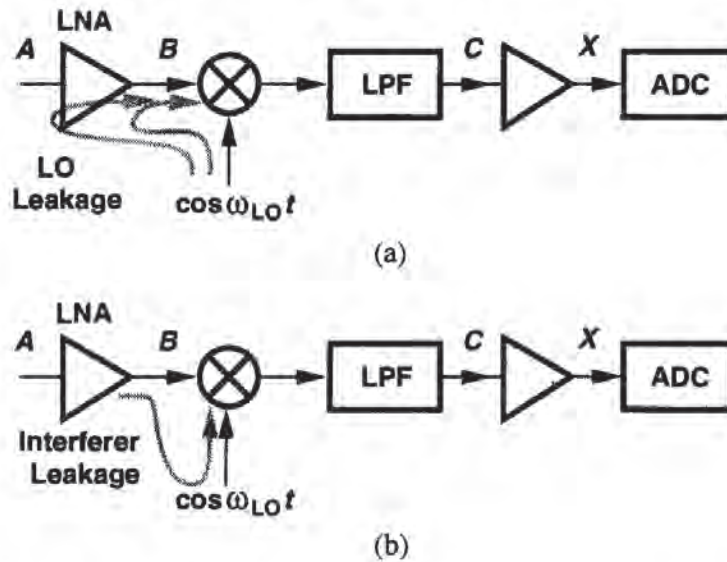


Figure 5.16 Self-mixing of (a) LO signal, (b) a strong interferer [5].

First, the isolation between the LO port and the inputs of the mixer and the LNA is not infinite; that is, a finite amount of feedthrough exists from the LO port to points A and B [Fig. 5.16(a)]. Called “LO leakage,” this effect arises from capacitive and substrate coupling and, if the LO signal is provided externally, bond wire coupling. The leakage signal appearing at the inputs of the LNA and the mixer is now *mixed* with the LO signal, thus producing a DC

component at point *C*. This phenomenon is called “self-mixing.” A similar effect occurs if a large interferer leaks from the LNA or mixer input to the LO port and is multiplied by itself [Fig. 5.16(b)].

Second, the total gain from the antenna to point *X* is typically around 80 to 100 dB so as to amplify the microvolt input signal to a level that can be digitized by a low-cost, low-power ADC. Of this gain, typically 25 to 30 dB is contributed by the LNA/mixer combination.

With the above observations, we can obtain a rough estimate of the offset resulting from self-mixing to appreciate the problem. Suppose in Fig. 5.16(a) the LO signal has a peak-to-peak swing of 0.63 V (≈ 0 dBm in a 50- Ω system) and experiences an attenuation of 60 dB as it couples to point *A*. If the gain of the LNA/mixer combination is 30 dB, then the offset produced at the output of the mixer is on the order of 10 mV. We also note that the desired signal level at this point can be as low as approximately $30 \mu\text{V}_{rms}$. Thus, if directly amplified by the remaining gain of 50 to 70 dB, the offset voltage saturates the following circuits, thereby prohibiting the amplification of the desired signal.

The problem of offset is exacerbated if self-mixing varies with time. This occurs when the LO signal leaks to the antenna and is radiated and subsequently reflected from moving objects back to the receiver. For example, when a car moves at a high speed, the reflections may change rapidly. Under these conditions, it may be difficult to distinguish the time-varying offset from the actual signal.

From the above discussion, we infer that homodyne receivers require some means of offset cancellation. While high-pass filtering may seem the natural solution here [4], two issues oppose the use of this technique. First, as explained in Chapter 3, the spectrum of many commonly used signals exhibits a peak at the zero frequency; that is, it contains substantial energy (information) near dc. For a 200-kHz channel, if the high-pass filter removes only the band from 0 to 20 Hz, the bit error rate rises to above 10^{-3} [5], indicating the need for a very low corner frequency. Second, in addition to demanding prohibitively large capacitors, this method also fails to track fast *variations* in the offset voltage, performing only a coarse cancellation.

The problem of offset can be alleviated using one of two techniques. First, the baseband signal in the *transmitter* can be encoded such that, after modulation and downconversion, it contains little energy near DC. Called “DC-free coding,” this is particularly suited to wideband channels, for example, in DECT, where a few kilohertz of the channel can be wasted with no significant drop in the data rate.

The second technique is to exploit the idle time intervals in digital wireless standards to carry out offset cancellation. Shown in Fig. 5.17 is an example, where a capacitor stores the offset *between* consecutive TDMA bursts, while introducing a virtually zero corner frequency during the reception of data. For a typical TDMA frame of a few milliseconds, offset cancellation is performed with sufficient frequency to take into account variations due to moving objects.

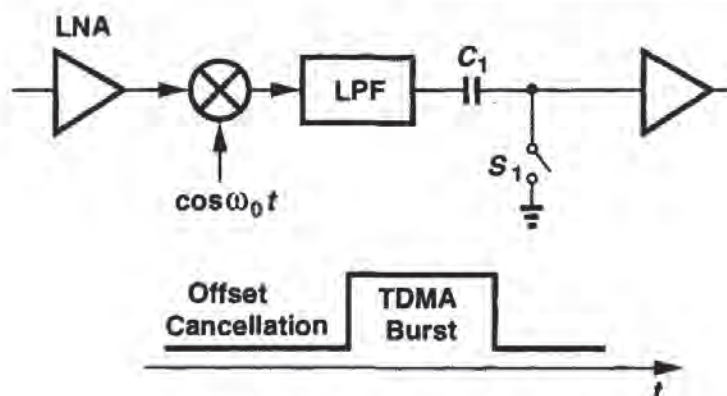


Figure 5.17 Simple offset cancellation in a TDMA system.

In practice, the thermal noise due to S_1 in Fig. 5.17 that is stored on the capacitor (kT/C noise) is significant, mandating large values for C_1 . For example, if a $1\text{-}\mu\text{V}_{rms}$ signal received at the antenna experiences a gain of 30 dB before offset cancellation, then C_1 must be at least 200 pF so that $\sqrt{kT/C}$ remains 15 dB below the signal level. For I and Q channels using differential circuits, this requires *four* large capacitors. However, if the offset is canceled after baseband preamplification [Fig. 5.15(b)], the value of the offset-storage capacitors can be reduced.

A general difficulty with offset cancellation in a receiver is that interferers may be stored along with offsets. This occurs because, as explained above, reflections of the LO signal from nearby objects must be included in offset cancellation, and hence the antenna cannot be disconnected (or “shorted”) during this period. While the timing of the actual signal (the TDMA burst) is well defined, interferers can appear any time. A possible approach to alleviating this issue is to sample the offset (and the interferer) several times and average the results.

We should also note that the problem of offset is much less severe in heterodyne architectures. Since the first LO frequency is not equal to the input carrier frequency, self-mixing may arise only for interferers [Fig. 5.16(b)], and dc offsets thus generated can be removed because the IF signal is far from zero frequency. Furthermore, in analog FM systems the second IF is nonzero and in digital modulation systems signal amplification and (partial) channel filtering at the first IF simplify the removal of the offset after the second downconversion.

I/Q Mismatch As shown in Fig. 5.14(b), for phase and frequency modulation schemes, a homodyne receiver must incorporate quadrature mixing. This requires shifting either the RF signal or the LO output by 90° (Fig. 5.18). As explained later in this chapter and in Chapter 7, shifting the RF signal generally entails severe noise-power-gain trade-offs, making it more desirable to use the topology in Fig. 5.18(b). In either case, the errors in the nominally 90° phase

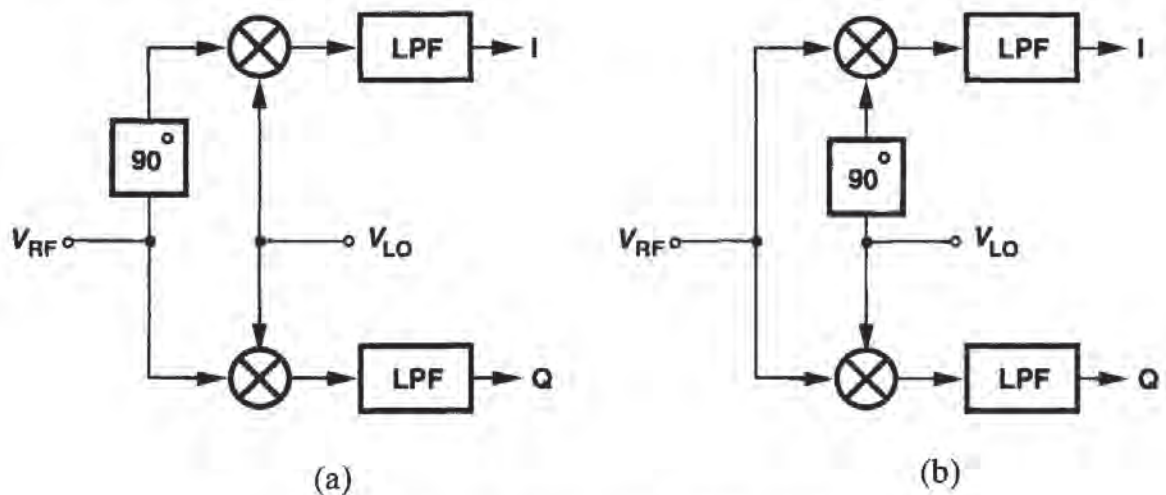


Figure 5.18 Quadrature generation in (a) RF path, (b) LO path.

shift, and mismatches between the amplitudes of the I and Q signals corrupt the downconverted signal constellation, thereby raising the bit error rate. Note that, as shown in Fig. 5.19, all sections in the I and Q paths contribute gain and phase error.

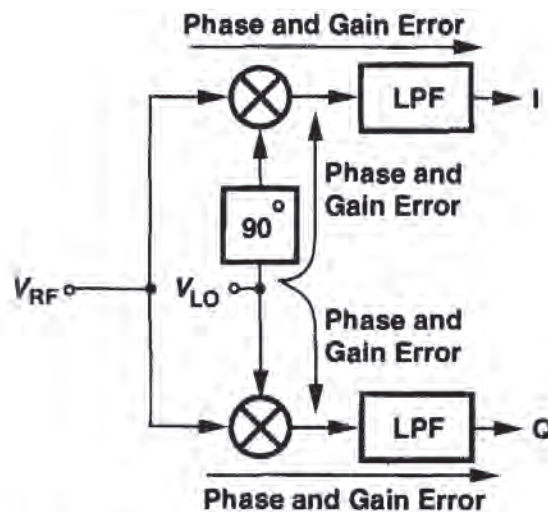


Figure 5.19 I/Q mismatch contributions by various stages.

To gain more insight into the effect of I/Q imbalance, suppose the received signal $x_{in}(t) = a \cos \omega_c t + b \sin \omega_c t$, where a and b are either -1 or $+1$. Now let us assume that the I and Q phases of the LO signal are

$$x_{LO,I}(t) = 2 \left(1 + \frac{\epsilon}{2} \right) \cos \left(\omega_c t + \frac{\theta}{2} \right) \quad (5.1)$$

$$x_{LO,Q}(t) = 2 \left(1 - \frac{\epsilon}{2} \right) \sin \left(\omega_c t - \frac{\theta}{2} \right), \quad (5.2)$$

where the factor 2 is included to simplify the results and ϵ and θ represent amplitude and phase errors, respectively. Multiplying $x_{in}(t)$ by the two LO phases and low-pass filtering the results, we obtain the following baseband signals:

$$x_{BB,I}(t) = a \left(1 + \frac{\epsilon}{2}\right) \cos \frac{\theta}{2} - b \left(1 + \frac{\epsilon}{2}\right) \sin \frac{\theta}{2} \quad (5.3)$$

$$x_{BB,Q}(t) = -a \left(1 - \frac{\epsilon}{2}\right) \sin \frac{\theta}{2} + b \left(1 - \frac{\epsilon}{2}\right) \cos \frac{\theta}{2}. \quad (5.4)$$

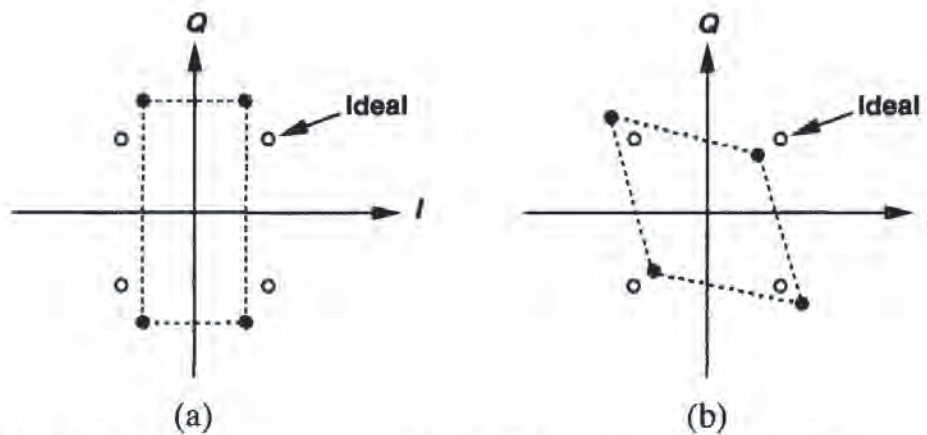


Figure 5.20 Effect of I/Q mismatch on QPSK signal constellation: (a) gain error, (b) phase error.

Fig. 5.20 shows the resulting signal constellation with finite ϵ or θ . This effect can be better seen by examining the downconverted QPSK signals in the time domain (Fig. 5.21). Gain error simply appears as a nonunity scale factor in the amplitude. Phase imbalance, on the other hand, corrupts each channel by a fraction of the data pulses in the other channel, in essence degrading the signal-to-noise ratio if the I and Q data streams are uncorrelated.

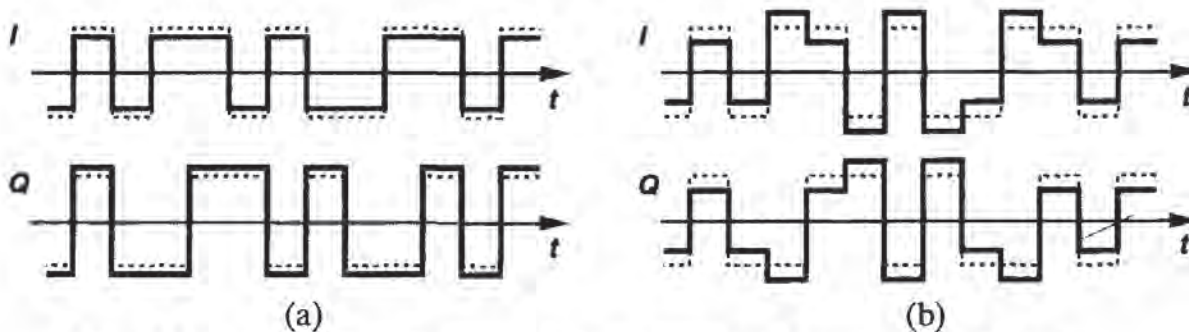


Figure 5.21 Effect of I/Q mismatch on a demodulated QPSK waveform: (a) gain error, (b) phase error.

In practice, it is desirable to maintain the amplitude mismatch below 1 dB and phase error below 5° , but these bounds depend on the type of modulation [7].

While heterodyne architectures may also employ I/Q downconversion in the last stage (Fig. 5.13), their mismatch requirements are much more relaxed. This is for two reasons. First, since the frequency at which I and Q phases are separated is about one to two orders of magnitude lower than that in homodyne counterparts, the two paths are much less sensitive to mismatches in parasitics. Also, in IC design, the lower frequency allows the use of large devices to improve the matching without excessive power dissipation. Second, in heterodyne receivers, the signal is amplified by approximately 50 to 60 dB before I/Q separation, requiring only one or two more stages afterwards. By contrast, each channel of a homodyne architecture incorporates several stages of gain and filtering, each of which contributes mismatches. We should also note that heterodyne topologies can perform the I/Q separation in the digital domain to avoid mismatch issues (Section 5.2.4) whereas homodyne receivers cannot.

The problem of I/Q mismatch has been an obstacle in discrete implementations, but it tends to improve as monolithic integration embraces more sections of homodyne architectures. Furthermore, since mismatches vary negligibly with time, signal processing techniques may be utilized to correct the points in the constellation.

Even-Order Distortion While only odd-order distortion has thus far been considered a source of interference (Chapter 2), even-order nonlinearity also becomes problematic in homodyne downconversion. Suppose, as illustrated in Fig. 5.22, two strong interferers close to the channel of interest experience a nonlinearity such as $y(t) = \alpha_1 x(t) + \alpha_2 x^2(t)$ in the LNA. If $x(t) = A_1 \cos \omega_1 t + A_2 \cos \omega_2 t$, then $y(t)$ contains a term, $\alpha_2 A_1 A_2 \cos(\omega_1 - \omega_2)t$, indicating that two high-frequency interferers generate a low-frequency beat in the presence of even-order distortion. Upon multiplication by $\cos \omega_{LO} t$ in an *ideal* mixer, such a term is translated to high frequencies and hence becomes unimportant. In reality, however, mixers exhibit a finite direct feedthrough from the RF input to the IF output. This is because, as explained in Chapter 6, mixers typically suffer from some asymmetry and their operation can be viewed as $v_{RF}(t)(a + A \cos \omega_{LO} t)$, where a is a constant. Thus, a fraction of $v_{RF}(t)$ appears at the output with *no* frequency translation. In typical differential mixers, the beat signal is attenuated by only 30 to 40 dB as it couples to the output.

The problem of even-order distortion can be studied from another point of view. Suppose in addition to phase or frequency modulation, the received signal exhibits some amplitude modulation as well. This could arise from filtering in the transmitter (Chapter 3) or disturbance and fading during propagation.

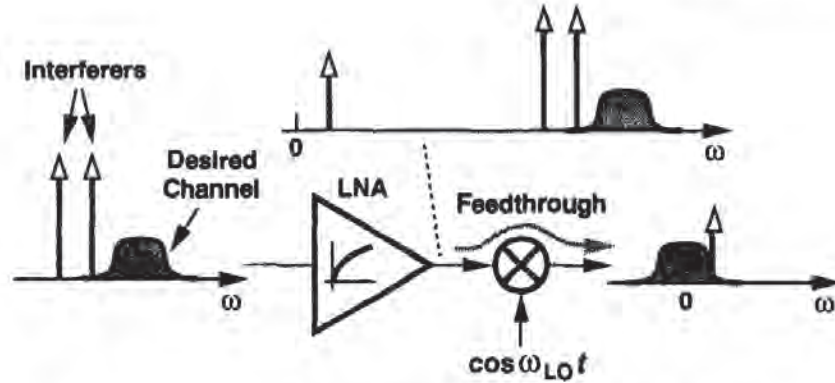


Figure 5.22 Effect of even-order distortion on interferers.

If $x_{in}(t) = (A + \epsilon \cos \omega_m t)(a \cos \omega_c t + b \sin \omega_c t)$, where $\epsilon \cos \omega_m t$ represents a low-frequency amplitude-modulating signal, then second-order distortion yields a term such as $(a^2 + b^2)A\epsilon \cos \omega_m t$. Since this term is the modulating signal translated to the baseband, it is often said that even-order distortion demodulates AM components. As mentioned above, this signal passes through the mixer with finite attenuation, thereby corrupting the downconverted signal of interest.

In our illustration here, we have assumed that only the LNA exhibits even-order distortion. In practice, the mixer RF port may also suffer from the same effect, requiring special attention in the design because the signals applied to the mixer are amplified by the LNA and can create significant distortion.

Second-order distortion can be characterized using the “second-order intercept point,” IP_2 . In a manner similar to the definition of IP_3 (Chapter 2), two equal-amplitude interferers are applied at the input and their low-frequency beat signal is observed at the output. Plotting the beat signal power versus the input power and extrapolating the results yield the IP_2 .

Can we use differential LNAs and mixers to suppress even-order distortion? Two issues must be considered here. First, the antenna and the duplexer filter are usually single-ended because they must operate with single-ended power amplifiers in the transmit path. Thus, a means of converting the received signal to differential form is necessary. The difficulty is that such means, for example, transformers, typically exhibit several decibels of loss at high frequencies, thus directly raising the overall noise figure. Second, if the LNA is designed as a differential circuit, it requires higher power dissipation than a single-ended counterpart to achieve a comparable noise figure.

Flicker Noise As mentioned earlier, the gain through the front-end LNA and mixer is typically around 30 dB, yielding baseband signal levels in the range of tens of microvolts. The input noise of the following stages, e.g., amplifiers and filters, is therefore still critical. In particular, since the downconverted spectrum extends to zero frequency, the $1/f$ noise of devices substantially cor-