RICHARD A. POISEL

INTRODUCTION TO COMMUNICATION ELECTRONIC WARFARE SYSTEMS

SECOND EDITION

For a listing of recent titles in the
Artech House Intelligence and Information Library,
turn to the back of this book.

Introduction to Communication Electronic Warfare Systems

Second Editon

Richard A. Poisel

ARTECH
HOUSE
BOSTON | LONDON
artechhouse.com

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress.

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library.

ISBN-13: 978-1-59693-452-8

Cover design by Igor Valdman

685 Canton Street
Norwood, MA 02062

10 9 8 7 6 5 4 3 2 1

To Debbie

Table of Contents

Preface to the Second Edition

This second edition was written with many of the same goals we had when writing the first edition with one major exception: the first edition was written with both engineering and operational audiences in mind. This turned out to be too broad an audience for a single book on this topic. As such, the second edition is much more focused on the engineering aspects of *electronic warfare* (EW), although operational topics are also discussed. The latter, in the case of the second edition, is for the edification of the engineering audience, however. Those areas are intended to add the "why" to the engineering discussions.

The first edition contained a very brief section on noise in EW systems—essentially only an introduction to the topic. It is such an important issue with EW systems that we thought it would be beneficial to expand on the topic and so there are substantial additions pertaining to the sources and affects of external and internal noise on EW system performance.

It is well known that considering the effects of signal fading can have substantial effects on the performance of communications systems, especially those that use digital modulations. The first edition considered fading effects only briefly. In the second edition, substantial additions pertaining to the effects of fading on jammer performance are included.

The simulated performance of a new EW architecture is presented. This architecture is based on the evolving concepts of electronically networked military forces. Borrowed from the world of computers, the jammers in this architecture carry the appellation "thin jammers" to emphasize the point that they have limited self-supporting target detection capability. Such jammers are distributed throughout a region, probably placed on the ground, with limited antenna elevation. They are controlled from a central facility, and a limited number of operators control them all.

Lastly, three chapters were removed in their entirety. These are: (1) receivers, (2) power amplifiers and exciters, and (3) antennas. These chapters were deleted from the second edition because we deemed them important enough topics to be the subjects of separate manuscripts in their own right.

So the organization of the second edition is as follows. There are three principal parts: (1) Introductory/motivational material in the first three chapters, (2) Part I – Basics, and (3) Part II – Performance. Part I goes over engineering material that many engineering students study in college; it is, however, focused on EW system issues. It is primarily intended for the reader that either did not have the material as a topic of study or had it long enough ago that a review is useful. Part II presents material on the performance of EW systems, generated for the most part by computer simulations.

After the introduction to the general topic of jamming, Chapter 2 presents some principles of *electronic support* (ES). Chapter 3 does the same for *electronic attack* (EA). Chapter 4 begins the basic principles section with a discussion of signal propagation. Chapter 5 continues with the expanded material on noise. Chapter 6 presents the fundamentals of communication technologies and is largely unchanged from the first edition. Following that, Chapter 7 presents the material on signal processing used in EW systems. It too is largely unchanged from the first edition with the exception of the addition of a discussion of detection of spread spectrum signal applications. Chapter 8 presents material on direction finding emitter geolocation techniques, while Chapter 9 contains the same topic but using quadratic processing. Chapter 10 starts the Part II – Performance section with a discussion of early entry ES. Chapter 11 presents the results of simulations on detection of frequency-hopping emitters. Chapter 12 contains the material on detection range limitations of EW systems. Chapter 13 is an extensive discussion of the effects of fading on jamming performance. Chapter 14 is essentially unchanged from the first edition with the exception that some of the material in that chapter in the first edition applies to the thin jammer architecture that is presented in Chapter 15, so it was extracted from Chapter 14.

As with any written technical work, there are bound to be errors that have crept into the material. I accept full responsibility for their presence and would appreciate any and all constructive feedback.

Preface to the First Edition

This book grew out of a need for an introductory level technical text on the basic aspects of communication electronic warfare (EW) systems. EW as a topical area consists of many aspects and probably the area treated least in detail is communication EW systems. Traditional books on EW signal collection and processing have concentrated on radar signals. This has been, in part, because of the closeness with which the governments of the world have held such information. After all, EW is a countermeasure technique targeted against, in this case, communication radio systems. The United States and other developed countries have produced radios to thwart such countermeasures as best they could. In the past that has not been very good and open discussions about how to attack radio communications ran counter to good sense, economically at least. That situation has changed of late, however. Effective radio systems to counter EW countermeasure methods have been developed and fielded. Therefore, the basic principles behind communication EW can now be discussed more openly.

When the author was new to the technical area, there was virtually nothing written specifically about communication EW systems. Everything had to be learned from first principles and experience. Needless to say, there were many mistakes made along the way. This book is an attempt to change that.

The book was written with two, not necessarily distinct, target audiences in mind. The first, but not necessarily the more important, consists of engineers first being introduced to the world of communication EW technologies and systems. There is a very real potential overlap of this group with the second group for which it was written: practicing EW military professionals. Some sections of the book contain derivations of equations, but those interested in the communication EW operational issues discussed can skip these sections.

For the technical audience, there is adequate depth in many areas for a concentrated and rewarding reading. The chapters on signal processing (Chapter 10), direction-finding position fixing (Chapter 11), and hyperbolic position fixing (Chapter 12), were written particularly with this audience in mind. The chapters that concentrate on operational issues (Chapters 1, 3, 5, 6, 7, 14, 15, 16, and 17) should appeal to this audience as motivational for studying the field in the first place. These chapters provide an understanding as to why there is an operational interest in EW as well as discuss some of the practical limitations.

For the audience with more of an operational interest, the more technical chapters can be skipped, although it would probably be better to just skip the math and focus on the operational sections of even the more technical chapters. These chapters also contain some important discussion of such issues. Chapters 1, 3, 5–7, and, in particular, 14–17 should appeal to those with operational interests. These chapters focus more on how EW systems are, or can be, employed. There are very real limitations on what such systems can do and many of these are presented.

Chapter 1

Communication Electronic Warfare Systems

1.1 Introduction

Just as the agrarian age, of necessity, gave way to the industrial age in the latter part of the 19th century, the industrial age gave way to the information age in the latter part of the 20th century. Pundits of modern armies comment about a revolution in military affairs, where attacking information systems (as well as protecting one's own) plays a substantial part in conflicts of the future. Thus is born military information warfare, sometimes referred to as command and control warfare, where protecting and attacking information and the systems that process it come into play.

1.2 Information Warfare

Information warfare (IW) is the appellation applied to conducting warfare-like actions against an adversary's information systems or protecting one's own information systems from such activities [1–3]. Attacking information systems can serve several purposes. It can deny the availability of key information at important junctures. It can provide false information causing an adversary to reach incorrect conclusions. It can degrade the confidence an adversarial decision maker has in his or her information bases also by providing false information. This is just a partial list of the effects of information warfare-like activities.

It is frequently stated that the first war in the Persian Gulf was the first information war [4, 5]. This is because of the extensive use of modern information technologies that affected the way the war was fought. This is not to say, however, that the *use* of information in war is a new concept. Information has always been important [6]. What is new are the methods implemented to use or affect information usage by an adversary. Modern information technologies permit such new affects.

IW can be applied to both nonmilitary as well as military situations. One commercial firm may conduct electronic espionage against a competitor in order to gain a competitive advantage. This is an example of commercial IW. Jamming the communications of an adversary when hostile activities are occurring is an example of military IW.

IW is generally (and incorrectly) construed to refer to attacks on computer networks. However, information warfare as a category can be interpreted to include just about everything that has to do with the use of information. This yields a very broad field and a fundamentally problematic taxonomy. Thus it has become accepted to refer to the more useful, but limiting, definition of attacks on computer networks. For our purposes, this is too limiting.

Command and control warfare (C2W) is IW applied in a military setting [7]. C2W is comprised of five so-called "pillars": (1) physical destruction of information systems, (2) psychological operations (PSYOPS), (3) deception, (4) operational security (OPSEC), and (5) electronic warfare (EW).

Physical destruction of information systems is pretty much self-explanatory. It can be accomplished in several ways; from hand emplacing explosive devices to missiles that home on radiated energy. It is, obviously, the extreme case in that, in general, hostilities must be under way, where the other forms of IW are subtler in their application.

PSYOPS has been used for years. It attacks the psyche of an adversary and it also can take many forms. Acoustic attacks can be used to try to convince an adversary to take some action. Pamphlets can be dropped from aircraft over adversarial forces to attempt to accomplish the same thing. In a sense, the simple unattended aerial systems (UASs) used in the first Gulf War were a form of PSYOPS in that just upon sight, the UAS took pictures of Iraqi solders surrendering *to the UAS*.

Deception involves actions taken to create the appearance of the existence of a situation that is not actually present. It can be as simple as placement of cardboard resemblances of vehicles in an area to give the impression of the presence of more forces than are actually present, to the transmission of radio communications in a region to give the same impression. The movement of coalition forces on the Saudi Arabian side of the Kuwait border created the appearance of an attack against Saddam Hussein's forces in Kuwait in the incorrect region [8].

OPSEC are procedures emplaced to ensure the protection of friendly information. Protecting classified information by proper utilization of security containers is an example. Taking care during telephone conversations not to give away what might be sensitive information is another.

Electronic warfare, the last pillar of IW, is the subject of the remainder of this book—it is discussed next.

1.3 Electronic Warfare

Throughout history, military warfare has been about measures to defeat one's enemy and countermeasures to those measures on the other side. In modern times, nowhere is this more evident than in information warfare, and in particular, communication EW. EW practices are always undergoing changes due to an adversary changing his or her capabilities based on a countermeasure developed by the other side. Theoretically this "measure/countermeasure" process could go on unbounded. In practice it does not, however, largely because of economics. Eventually things get too expensive. This is true in electronic warfare just as it is true in other areas.

This book is about that part of IW that attacks information systems by withdrawing from, or imparting energy into, communication systems so that the intended transport of information is either intercepted, denied, or both. As indicated above, that part of IW is called communication EW. More specifically, the systems that are intended to attack communication systems are discussed. By simple extension, these systems can sometimes be used to "screen" friendly communications from similar actions taken by an adversary so they can also be used in an information protection role.

Most of the more important salient attributes of these systems are covered herein, although no claim is made that all facets are included. There will be derivations of equations. Such derivations are included in a few places to indicate that most of the aspects of designing EW systems are based on fundamentally sound first principles, and these cases illustrate some of these first principles.

EW can be, and is, applied to signals that are virtually anywhere in the entire frequency spectrum. For our purposes we will limit the discussions to what is normally referred to as the radio frequency part of the spectrum. This starts about 500 kHz and extends into the 100's of GHz. This specifically excludes audio frequency measures at the low end, and infrared and electro-optics at the high end.

It is generally accepted that EW has three distinct components: (1) *electronic support* (ES), (2) *electronic attack* (EA), and (3) *electronic protect* (EP) [9]. Electronic support is comprised of those measures taken to collect information about an adversary by intercepting radiated emissions. Electronic attack refers to attempting to deny an adversary access to his or her information by radiating energy into their receivers. Electronic protect involves activities undertaken to prevent an adversary from successfully conducting ES or EA on friendly forces. The U.S. military joint doctrine on EW is illustrated in Figure 1.1 which is extracted from joint publication 3-51 (2000) [10].

1.3.1 Electronic Support

Electronic support attempts to ascertain information about an adversary by intercepting radiated energy. This radiated energy can be from any type of

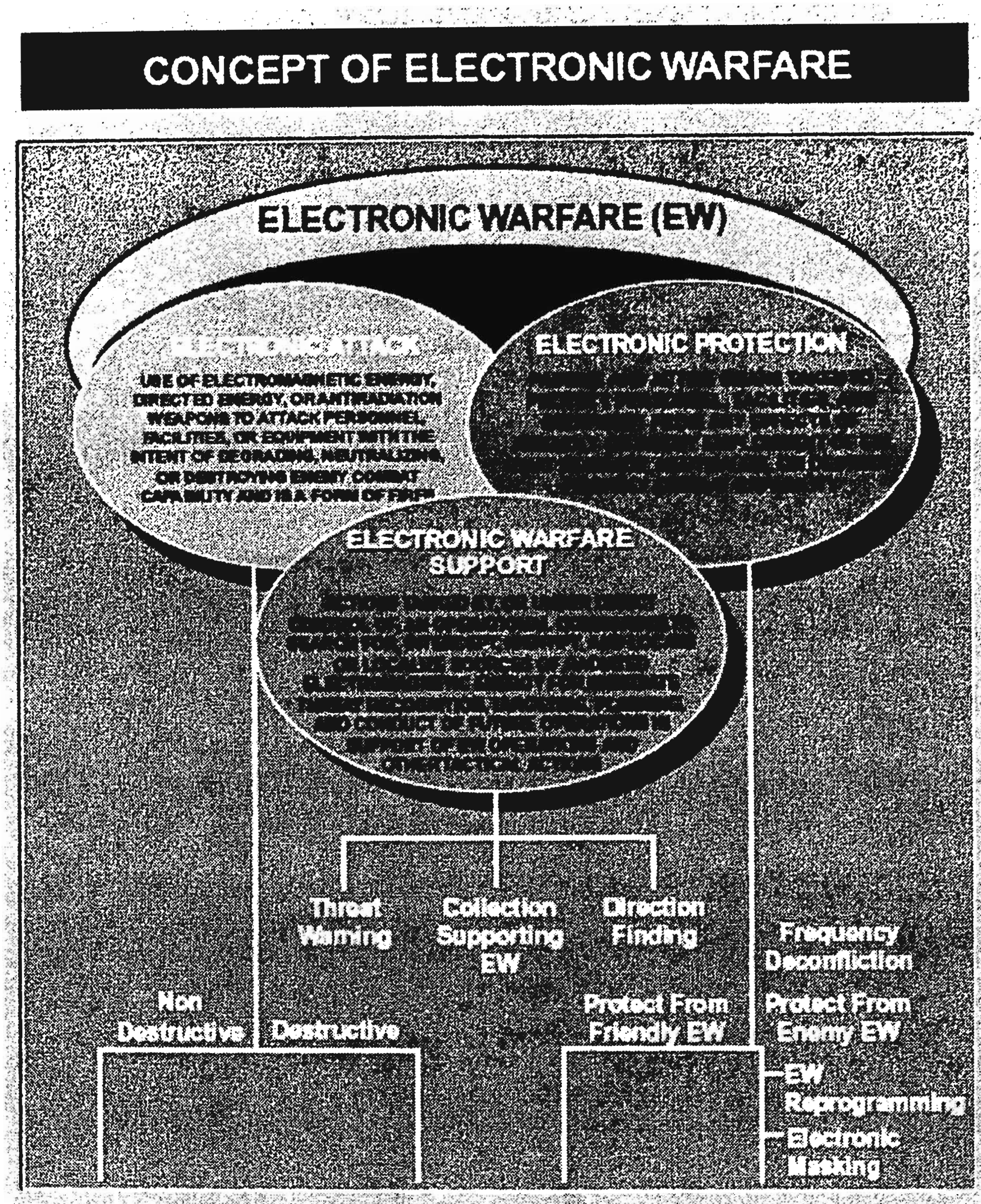

CONCEPT OF ELECTRONIC WARFARE
ELECTRONIC WARFARE (EW)
ELECTRONIC ATTACK
USE OF ELECTROMAGNETIC ENERGY, DIRECTED ENERGY, OR ANTIRADIATION WEAPONS TO ATTACK PERSONNEL, FACILITIES, OR EQUIPMENT WITH THE INTENT OF DEGRADING, NEUTRALIZING, OR DESTROYING ENEMY COMBAT CAPABILITY AND IS A FORM OF FIRES
ELECTRONIC PROTECTION
ELECTRONIC WARFARE SUPPORT
Threat Warning
Collection Supporting EW
Direction Finding
Frequency Deconfliction
Non Destructive
Destructive
Protect From Friendly EW
Protect From Enemy EW
EW Reprogramming
Electronic Masking

Figure 1.1 U.S. military concept of EW from joint publication JP 3-51. (*Source*: [10].)

transmitter such as those in communication networks. It could also be from radars, telemetry transmitters, or unintended radiation as from computer clocks.

Some important information can be gleaned from just the measurement of a few external parameters associated with a transmission. The frequency of operation, the modulation type, the bit rate, or the geolocation of the transmitter are examples of such external parameters. In some cases it is also possible to intercept the internals of a transmission. When this occurs it is normally termed SIGINT, where the goal is to generate intelligence products about an adversary. ES is usually restricted to collection of external parameters.

1.3.2 Electronic Attack

When radiated energy from a friendly source is used to deny an adversary access to his or her information, it is referred to as electronic attack. Again the adversarial radiated energy could be from communication transmitters, radars, telemetry transmitters, and so forth. It is important to remember, however, that in electronic attack it is the receiver that is the target of the attack. Examples of electronic attack include radio jamming and radar deception.

1.3.3 Electronic Protect

It is frequently desired to take measures to disallow an adversary to conduct ES and EA against friendly forces. As dominant battlespace knowledge[1] [11] becomes more and more important in the era of information warfare, more reliance on accurate, timely, and thorough information is necessary. Therefore it becomes more important to protect friendly information from manipulation by an adversary. It is also important to deny the availability of friendly information to those adversaries.

Many disciplines can be grouped within EP. Emission control (EMCON) is perhaps one of the simplest forms. Through EMCON, the use of friendly transmissions is limited or precluded for a certain period of time, usually at critical junctures. Simply the presence of such emissions can provide information to an adversary as to friendly force sizes and, perhaps, intentions. EMCON prevents an adversary from intercepting and identifying the operating frequency of a friendly communication network, thus protecting the frequency information from being available to the adversary. Another form of providing such protection is low probability of intercept (spread spectrum) communications (discussed in Chapter 6).

While EMCON in general is part of the overall communication and operations planning process, and is normally accomplished manually, automation of the

[1] *Dominant battlespace knowledge* (DBK) is knowing more accurate information about the battlespace at a particular point in time than the adversary. See [10] for a thorough discussion of DBK.

EMCON concept for optimum control of communications, especially sensor networks, has recently been described [12].

Screen jamming is a form of EP. This is when a jammer is placed between friendly communication nets and an adversary's SIGINT systems to prevent the latter from intercepting the communications. Screen jamming is a useful tool when attempting to break contact to restructure the nature of hostilities, for example.

Encrypting communication nets is another form of EP. It prevents an adversary from gleaning information from the intercept of communication transmissions. The ready availability of practical and effective encryption algorithms to all countries is causing more and more communications of all types to be encrypted.

EP is a category in a much larger area called information assurance. In general, assuring the integrity and security of information is a complex task in a complex world. The U.S. military is just beginning to get a handle on the enormity of the problem [13].

This book is primarily about ES and EA against communication targets. Many of the technologies applied to ES and EA discussed, however, are applicable to EP. Little more will be said specifically about EP herein.

1.4 Electronic Support

ES refers to those measures taken to gather information from radio frequency emissions, by noncooperative intercept of those emissions. When the emissions are from a communications system, then it is known as communications ES. Herein, the communication will be assumed and communication ES will be referred to as simply ES. The term noncooperative in this sense means that the communicator and interceptor are not cooperating—the communicator does not want the interceptor to be successful.

ES is the collection of communication signals for the purposes of gleaning information from them. This information can be in the form of intelligence or combat information. The difference between these two is what the information is used for as opposed to the collection means. Another difference might be the type of information collected—signal externals versus internals, for example. Generation of intelligence usually requires access to the internals of a signal, a human analyst, and a relatively extensive amount of time. On the other hand, combat information is information that is readily apparent from the data and does not require extensive analysis—usually no human analysis at all. The externals of a signal are usually all that are used to generate combat information. Externals are signal parameters such as frequency of operation, baud rate, location, etc. Combat information can be used directly to support ongoing operations, such as real-time targeting, for example. Combat information, of course, can be, and is, also used for intelligence generation.

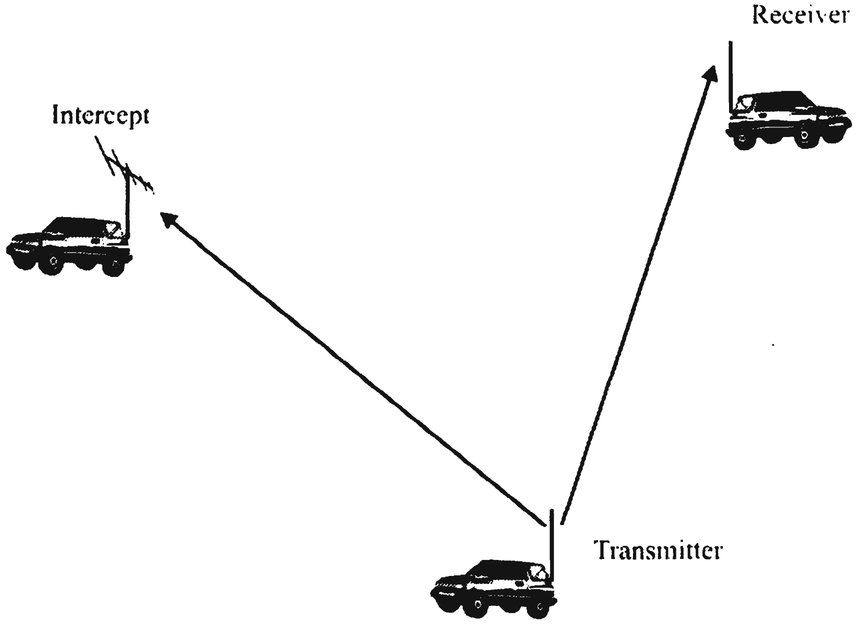


Figure 1.2 Notional scenario for ground-based electronic support. Copyright © 2001 Artech House and licensors. All rights reserved.

A ground-based ES scenario is depicted in Figure 1.2. The transmitter is communicating with a receiver, while the ES system is trying to intercept that transmission. The fundamental distinction between ES and the normal reception of communication signals by the intended receiver is that the former is noncooperative while the latter facilitates communication between the transmitter and the receiver. That is, the transmitting entity is not trying to communicate with the ES system but it is with the receiver. In older forms of communication, analog AM and FM for example, this distinction was not so important. The reception of signals by either the receiver or the ES system depended mostly on the geographical relationship between the entities involved. With modern digital communications, however, this distinction is significantly more important.

Cooperating communication systems can adjust for imperfections in the communication channel between them. An example of this is evident when computer modems are first interconnected for telephone Internet communications. The tones and rushing noises heard at the beginning are the modems doing just that—determining the quality of the channel that interconnects them (normally, but not always, these days one or more telephone lines) and setting the communication speed depending on these measurements. A noncooperative ES receiver does not share this luxury. The transmitter not only does not try to measure the channel between the transmitter and the ES system, but also normally would not even want the ES system there at all. This simple distinction between cooperative and noncooperative communication situations to a large extent determines the complexity of the ES system.

Classical ES for communication signals usually entails determination of the geolocation of the emitting entity, determining its frequency of transmission, if appropriate, determining the type of signal it is (e.g., analog vs. digital), and, if possible, the type of weapon system the signal is associated with, such as artillery, rocket forces, etc. Such information is useful in determining the *electronic order of battle* (EOB), which refers to the entities on the battlefield and what kind of RF electronic systems they possess.

1.4.1 Low Probability of Detection/Interception/Exploitation

These terms apply to ways of processing electromagnetic (EM) signals so that either it is difficult to know if the signal is present or, if it is detected, the information contained in it is difficult to extract [14, 15].

Low probability of detection, or LPD, attempts to "hide" the presence of signals. One technique puts the signal below the level of the noise, so that the signal cannot be differentiated from the noise upon a casual look at the spectrum. This technique is called *direct sequence spread spectrum* (DSSS). Another technique jumps the signal around in frequency so that a fixed tuned receiver will rarely see the signal. This technique is referred to as *frequency-hopping spread spectrum* (FHSS).

If the nature of the communication system is such that it is difficult to establish LPD, then it may be more desirable to yield to the possibility that the signal may be detected, but that once it is, it is difficult to extract the information contained in it. *Low probability of interception*, or LPI, and *low probability of exploitation*, or LPE, are the terms used in this case. FHSS is one example of this technique.

Most EM signals that the reader may be familiar with, such as TV stations or FM and AM radio stations, are fixed in frequency. Most practical low probability schemes take such narrowband signals and spread them out so they occupy much larger portions of the electromagnetic spectrum than they would otherwise. One technique that doesn't spread the signal out in frequency but instead stretches it out in time is called "time hopping." This technique is not used as much as the other LPI techniques for precluding detection. Time hopping naturally occurs, however, for military tactical communications when it is not known a priori when an entity is going to communicate. Such communications are frequently referred to as *push-to-talk* (PTT).

Frequency hopping changes what's called the carrier frequency often, say 100 or more times per second. This is akin to shifting the station that your car radio is tuned to 100 times per second. Even if you knew where the next frequency was, and one of the features of frequency hopping is that unintended receivers do not, it would be difficult to tune your radio fast enough to keep up. Of course with modern digital receivers, the switching speed isn't a problem but not knowing the next frequency is.

The direct sequence spread spectrum technique mentioned above for generating low probability signals (the focus here is on LPD) is a little more difficult to understand. In this technique, a special signal called the chip sequence is added (modulo two, 0 + 0 = 0, 0 + 1 = 1, 1 + 0 = 1, 1 + 1 = 0) to the information signal (that has already been converted to digital) before it is broadcast. The chip sequence is a digital signal that changes states at a much higher rate than the information signal. The net effect is to have the resultant signal much broader in frequency extent, and at any particular frequency much lower in amplitude, than the information signal would otherwise occupy (if located at that frequency). It is much lower in amplitude at any single frequency since the same signal power is present, but it is spread across a much broader frequency range. Such a low amplitude signal can be difficult to detect if you do not know it is there.

A technique that has nothing to do with technology but is an effective LPD/LPI/LPE technique is called EMission CONtrol, or EMCON, mentioned above. This is using the technique of radio silence during certain portions of an operation to preclude radio signals from being detected, which in some cases would/could indicate that an operation is occurring. To be detected while employing EMCON would require a sensor other than a SIGINT sensor.

Collection of wideband signals such as these requires receiving equipment that is also wideband. Unfortunately, it is a law of physics that the larger the bandwidth of the receiving equipment the more background noise enters the receiver along with any desired signals. Nevertheless, there are some techniques that have been developed, including compressive receivers and digital receivers, which help to minimize this effect.

1.4.2 Future Communication Environments

As discussed in Poisel and Hogler [16], future communications of concern to the U.S. military forces will be largely based on commercial technology developments. The principal reasons for this are diminishing military budgets and increasing sophistication of commercial communications means around the world. Communication is one of the cornerstones of the information age, and will be vitally important to everyone living in it eventually. Therefore, the private sector is expected to invest heavily in the development of advanced communication means, precluding the necessity of military investments in communication technology—other than, of course, to adapt it to military applications and purchase it. Such adaptation will range from no modifications at all to ruggedizing the housing and electronics inside. The underlying signaling technology will not often be modified.

1.4.3 Electronic Support Summary

Interception of communication signals on the battlefield, herein collectively called ES, will continue to be an important source of information. If the communication signals are not encrypted and then in some cases the intent of an adversary can be inferred. Even if they are encrypted, however, information can still be extracted such as locating where the signals are coming from, the volume of signals, and other external parameters can be used.

Modern communication devices have gone to great lengths to prevent their interception. These techniques are referred to as low probability of intercept, low probability of detection, and low probability of exploitation. Wideband receiving systems must be used against such signals and the ranges of these systems is frequently limited because of the amount of noise associated with wide bandwidths. Compressive and other forms of analog receivers, as well as digital receivers, have been designed to be used against these signals.

The accuracy at which these signals can be geographically located depends strongly on the *signal-to-noise ratio* (SNR) of the signal at the receiver. Low SNR signals are more difficult to locate accurately than those with large SNRs. The accuracy normally increases as the inverse square root of the number of samples of the signals that are taken, as well as inversely to the integration time and measurement bandwidth.

1.5 Electronic Attack

Denying an adversary the effective use of his of her communication systems is the goal of communication EA. Any electronic means of attacking such targets, short of lethal weapons, falls under the guise of EA. The scenario shown in Figure 1.3 summarizes a situation for EA on communication nets, in a ground-based setting. The transmitter is attempting to transport information to the receiver while the jammer is attempting to deny that transport. This denial is in the form of placing more jamming energy into the receiver than the transmitter's information signal. The particular communication signal involved does not matter; however, the effectiveness of the jammer does depend on the signal type.

Conducting EA from standoff EW systems is complicated by the fact that frequently fratricide occurs—that is, interfering with friendly communication systems. It is therefore sometimes more fruitful to place EW systems closer to the communication nets that are being jammed than to friendly nets. Delivery mechanisms for such systems could include artillery shells—the 155 mm howitzer shell, for example, can go on the order of 20 km. Another delivery mechanism could include emplacing them by hand, or flying them in a UAS. In the latter case the EW systems could either be used on the UAS or they could be dispensed. Since the signal propagation range is greater from the UAS, one would expect better performance from this configuration, although one could encounter the fratricide problem again. Directional antennas on UAS platforms could help the fratricide problem although this solution wouldn't be much help in a nonlinear battlespace.

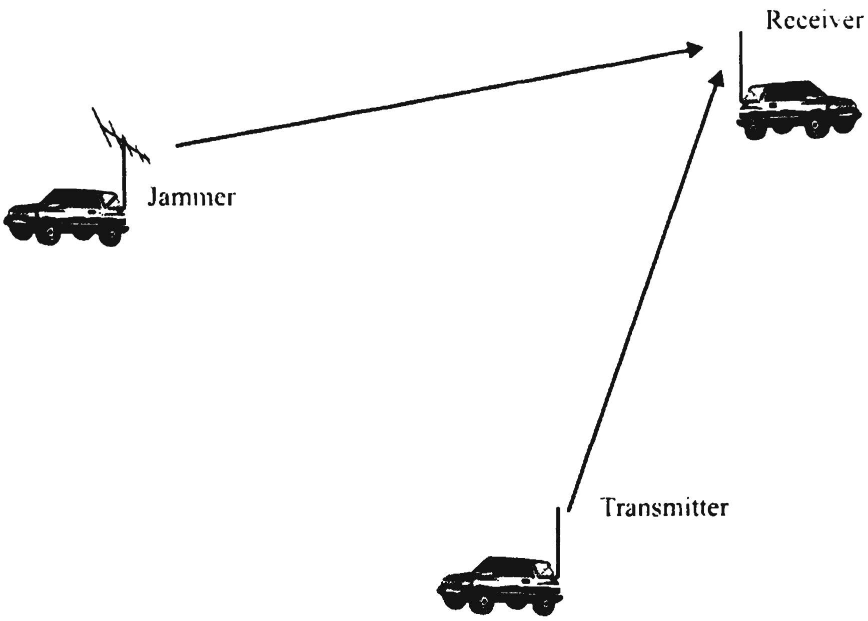


Figure 1.3 Notional scenario for ground-based electronic attack. Copyright © Artech House and licensors 2004. All rights reserved.

1.5.1 Electronic Attack Summary

Communication EA is employed in all forms of combat situations: ground to ground, ground to air, air to ground, and air to air. It is used to deny or degrade an adversary's ability to command and control forces. The effects of communication EA are temporary—in most cases they last as long as the jamming signal is present. Digital communication signals can be affected easier than analog signals, because a threshold *bit error rate* (BER) is required in order for digital communications to work. Typically this is about 10^{-2} or better for machine-to-machine communications. Thus causing problems on one bit out of 100 or more should be sufficient to deny machine-to-machine digital communications. Encrypted communications require resynchronization when synchronization is lost due to jamming or other reasons.

1.6 Introduction Summary

This book provides an introduction to EW. It presents the basic information about the topic intended for the newcomer to the field; albeit the material included has a significant engineering bent to it.

The remainder of this introductory chapter contains a discussion of the blocks in a block diagram of a generic communications EW system. It provides an introduction to some of the nomenclature and terminology. The second chapter provides an introduction to ES while the third provides the same for EA. These three chapters make up the introduction to communication EW and are largely motivational in nature. The next four chapters present detail on basic concepts associated with EW. Chapter 4 provides some background on electromagnetic wave propagation, with an emphasis on the characteristics common to lower frequency ranges where typical military command and control communication signals are found. Chapter 5 provides considerable detail on the noise and interference encountered by EW systems. This noise includes both sources internal to the system as well as noise generated by sources external to the system. An introduction to modern radio communication technologies is presented next in Chapter 6. To wrap up the basic information section, Chapter 7 presents aspects of signal processing common to many areas in EW systems.

The last eight chapters present EW-specific material, with an emphasis on performance parameters. One of the most important functions provided by EW systems is the geolocation of targets. Chapters 8 and 9 provide introductions to the two most common ways to accomplish this. Chapter 8 provides an introduction to direction finding (DF), where the angle of arrival of a signal is measured at two or more receiving sites. The geolocation is then computed by determining the intersection of these *lines of position* (LOPs). Chapter 9 contains material on quadratic target location systems, where the *time of arrival* (TOA), *time difference of arrival* (TDOA) between two or more systems, and the differential frequency, otherwise known as the *frequency difference of arrival* (FDOA), are measured. Under perfect conditions the emitting target lies on the resulting isochrones.

Chapter 10 continues with a presentation of the results of a simulation analysis for EW systems with limited capability, suitable for deployment with early entry forces to provide a limited organic capability. Chapter 11 provides a similar analysis for the capability to collect FHSS networks. Chapter 12 discusses the intercept range of typical EW systems, while Chapter 13 analyzes jamming performance in fading channels. Chapters 14 and 15 describe the jamming performance of some particular EA architectures. Chapter 14 includes some ground and air jammer configurations. Finally, the last chapter, Chapter 15, describes jamming performance of thin jammers, which consist of very limited capability jammers distributed throughout an area. Such jammers would be controlled from a central site and controlled over the network dispersed throughout the area.

There are three appendices at the end of the book where some material is presented that does not fit well with the flow of the rest of the book, yet was deemed important enough to include. The first of these is a tutorial on probability. The second contains the network configurations that are used in all the simulation analyses discussed. The third appendix provides an introduction to system engineering.

1.7 Typical EW System Configuration

Presented in this section is a configuration for a "typical" communications EW system. It will be used throughout this book so that the usage of each component can be illustrated. It certainly is not the only configuration available for RF EW systems, the design of which depends on the particular application. A block diagram for the typical EW system is shown in Figure 1.4. Some of these components may not exist in every system, and some systems may contain others. If the EW system were for ES only, the exciter, *power amplifier* (PA), filters, and EA antenna would not be present, for example. In this figure, the lighter lines represent radio frequency (RF) or other types of analog signals whereas the darker lines represent digital control signals.

Modern communications used by military or nonmilitary forces changed over the last quarter of the 20th century. World War II saw the widespread use of single-channel HF, VHF, and UHF radios. After WW II, the higher frequencies started to be exploited for communication purposes, satellite communications for example. At the end of the century, cellular phones and pagers were proliferating at rapid rates. Initially these devices operated at relatively (by today's standards) low frequencies—around 450 MHz. That spectrum quickly became overcrowded and it was obvious that with the demand for these telecommunication services, additional spectrum was required. Such spectrum was subsequently added around 900 MHz. As it stands today, spectrum around 1.8 GHz and in the 2–3 GHz band have been added.

Classical engineering practices for designing communication systems (and therefore communication EW systems as well), only apply to relatively low frequencies. In addition the size of electronic circuits is continually shrinking. Once the signal propagation path extends to a significant fraction of its wavelength, these traditional design techniques no longer apply. An engineering rule of thumb for this distance is $\lambda/10$. Since a signal at 3 GHz has a wavelength (λ) that is 10 cm, for a 3 GHz signal, this limit is 1 cm. Therefore for propagation paths longer than 1 cm, at 3 GHz, design techniques for high frequencies must be employed. A practical frequency limit is about 150 MHz, above which low-frequency design techniques should not be used. Because of this, the implications and impacts of high and microwave frequency signals are included.

1.7.1 System Control

One or more computers typically exercise the control of such a system. If there is only one, then the control is centralized. If there is more than one, then the control can be either centralized or distributed, depending on the chosen architecture.

The solid dark line indicates a system control bus, which may actually consist of more than one. Typical busses would be Mil Std 1553, VME, VXI, IEEE 1394

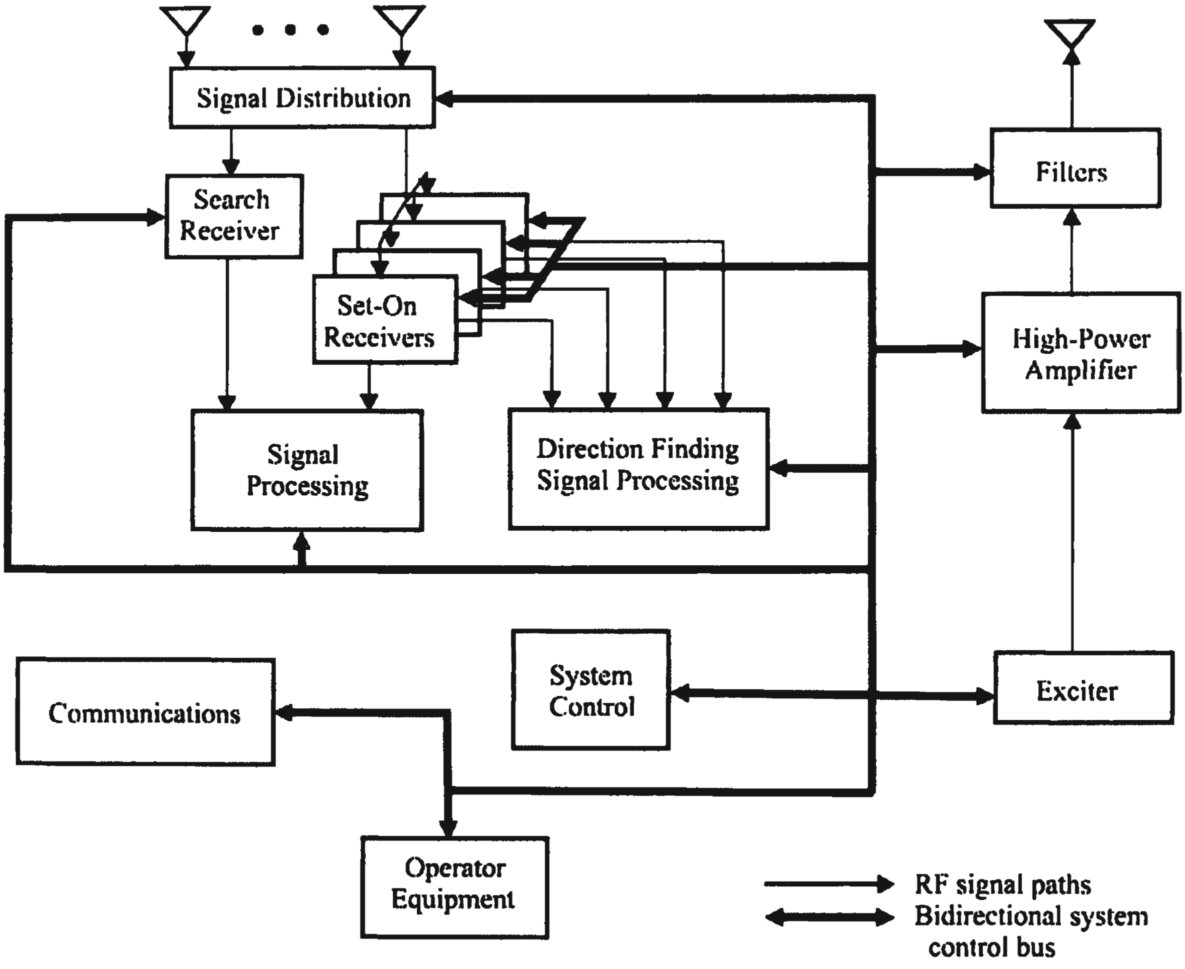


Figure 1.4 Block diagram for a typical EW system.

(Firewire), IEEE 802.3 (Ethernet), and ANSI X3T9.5 fiber distributed data interface (FDDI). The lighter lines are intended to indicate signal paths, whether RF or lower in frequency. If there are operators present that interface with the system, then there are usually workstations connected to the system control for that purpose.

1.7.2 Antennas

Antennas are used to extract the EM energy from the propagation medium. For most cases of interest herein the propagation medium is the atmosphere surrounding the Earth (the air). Antennas convert the EM energy into electrical signals, which is a form for the energy that the rest of the system can use.

Antennas are also used in the reverse fashion for EA applications. They convert electrical energy into EM energy that can be propagated through the atmosphere.

Typically there is more than one antenna performing several tasks associated with an RF EW system. Intercept and direction-finding antennas are frequently the same. Intercept and EA antennas also are frequently the same. The specific antenna configuration used depends on the application. Some exhibit gain over other types, thereby increasing the range of operation of the EW system. In some situations, using antennas with gain is troublesome due to mobility requirements, for example. Generally the higher the antenna is above the ground the farther is the range of operation. Thus airborne EW systems typically have a larger range than ground-based systems (for VHF and above frequencies anyway).

1.7.3 Signal Distribution

Signals from antennas frequently must go to more than one place in an EW system. In this example, the signals go to more than one receiver. In order to maintain the proper impedance the signal splitting must be done carefully. With improper impedance matching, less than maximum power transfer will result, and the signals could experience distortion that they otherwise would not.

Thus signal distribution is accomplished normally at the output of the antennas and before the receivers. Typically the antenna impedance is converted to the cable impedance, say 50 Ω. This cable is run from the antenna to the signal distribution and from there to the receivers. A signal splitter in the signal distribution would have an input impedance of 50 Ω and two or more output ports that have 50 Ω output impedance.

1.7.4 Search Receiver

A separate book is in preparation that discusses receivers in depth. Suffice it here to say that there is usually a function to be performed in an RF EW system that searches the frequency spectrum looking for signals of interest. A search receiver accomplishes this function. This receiver typically scans through the spectrum of interest, looking for energy. When it finds energy, measurements are made to characterize it. Those measurements can be performed as part of the search receiver, or by using a separate set-on receiver.

1.7.5 Set-On Receivers

A set-on receiver is used for relatively long-term analysis of signals detected by other means. There is usually more than one of these. They are tuned to a frequency either by an operator or automatically based on energy detected by the search receiver. They may, in fact, be nothing more than channelized filters using the search receiver as the RF portion. The outputs of these receivers are used to measure parameters of signals or, in the case of analog communications, for example, are for the operators to listen to.

Modern set-on receivers are typically digitally controlled where the digital control signal can change any of the parameters of the receiver, such as frequency and IF bandwidth.

1.7.6 Signal Processing

Much more will be said about signal processing later. This function comes in many forms, depending on what is trying to be accomplished. Typical functions include detection of the presence of energy at a particular frequency and within a specified bandwidth, determination of the modulation on a signal, measuring the baud rate of a digital communication signal, and so forth.

1.7.7 Direction-Finding Signal Processing

This signal processing function is broken out because of its importance in many applications. Locating the source of emissions of communication signals is one of the fundamental applications of RF EW systems. This is usually, but not always, accomplished by triangulation where the direction of arrival of a signal at two or more EW systems is measured. The point where the lines corresponding to these arrival angles intersect is the calculated location of the source of the signal. If all of the lines of bearing do not intersect at a single point (and they rarely do for more than two systems) then the emitter location is calculated according to some algorithm to be the centroid of the lines of bearing.

Therefore accurate measurement of the arrival angle is an important function. Various factors enter these processes that make it difficult to measure the angles accurately. Not the least of these is RF noise, caused by the stars in the sky. Interference is another source of noise. This is caused by man-made sources of energy that share the same frequency bandwidth as the signal of interest.

1.7.8 Exciter

If the EW system has an EA function to perform, then it typically will contain an exciter, high power amplifier, filters, and antenna as shown. An exciter is essentially an RF signal generator, with the capability to modulate the signals generated. The modulation can take many forms, depending on the target signal, but the one used most frequently for communication signals is simply random noise frequency modulated onto the carrier. This signal essentially simply raises the noise level at the target receiver, thus decreasing the SNR. The SNR is a fundamental parameter in communication theory and has been heavily studied. Therefore accurate prediction of degradation of performance due to jamming a signal can be made. Other forms of modulation that can be used include tone jamming, sometimes effective against *frequency shift key* (FSK) signals.

1.7.9 Power Amplifier

The high-power amplifier for electronic attack purposes amplifies the signal(s) from the exciter. The PA converts the relatively weak (typically 0 dBm, or 1 milliWatt) signal from the exciter into stronger signals for transmission. For communication EA applications, the signal level sent to the antenna is typically 1 kW or so.

Some of the bigger issues associated with power amplifiers include conversion efficiency (fraction of prime power converted to radiated power), frequency coverage (broad bandwidths are difficult to achieve with a single amplifier), heat removal (related to efficiency and reliability), and spectral purity to prevent friendly signal fratricide.

1.7.10 Filters

In order to avoid RF fratricide of friendly receivers, filters on the output of the amplifier are frequently needed, since perfect amplifiers with no spurious responses have yet to be invented. The filters limit the out-of-channel (undesirable) energy that the system emits. These filters must be constructed so they can handle the power levels associated with the output of the PA. Ideally they would introduce no loss of in-band signal power as well.

These filters must be tunable. Typically a jammer will radiate energy at a single frequency and over a limited frequency extent (bandwidth). The filters must be tuned to that frequency if they're to function properly. In some applications, such as EA against frequency-hopping targets, where the frequency is changing rapidly, this tuning must be performed even faster than the targets change frequency.

Obviously if there is more than one signal to be jammed by the communication EW system, there is more than one filter required, although it may not be necessary to provide more than one PA. The PA itself is wideband, capable of amplification over a broad frequency range. The exciter determines the frequency at which the jamming signal is located, and the filter keeps the signal within a specified bandwidth. Therefore an exciter and filter are required for each frequency, but only one PA is required.

The output filters may be required to have tunable bandwidths as well. Some modes of EW require (relatively) broadband noise waveforms. The output filters must be capable of passing these waveforms while blocking undesired signals at frequencies outside the band of interest.

1.7.11 Communications

The communication subsystem can be comprised of several types of capabilities. It is the means for command and control of the system as well as the means to send external data to it (tasking), and the mean to deliver a product out of it (reporting). If the system is remotely controlled, then this subsystem is the means to exercise that control.

1.8 Concluding Remarks

Contained in this introduction are brief descriptions and discussions about most of the more significant issues associated with radio frequency electronic warfare systems. Both electronic support and electronic attack are introduced. The remainder of this book will discuss these subsystems in more detail. This introduction is intended to describe the EW system as a whole before the parts are described in detail.

EW systems constitute the ability to intercept as well as take countermeasures against radio frequency signals. This information can be used to glean EOB data about an adversary as well as information on when and how to attack an adversary's command and control and sensor capability. As with any weapon system, the utility of an EW system depends on how it is employed and the situation to which it is applied. In some scenarios, an EW capability may be useless. In other scenarios it might be vital. An example of the latter situation occurred in the war in the Gulf in 1991. The Iraqi military was so concerned about the Allied's ability to employ communication EW systems that communication with their forward-deployed troops was almost nonexistent. Whether the EW systems that were employed were effective at performing the EW mission or not is irrelevant. They accomplished their goal simply by reputation.

References

[1] Libicki, M. C., "What Is Information Warfare," The Center for Advanced Concepts and Technology, The National Defense University, Washington, D.C., August 1995.

[2] Schleher, D. C., *Electronic Warfare in the Information Age*, Boston: Artech House, 1999.

[3] Waltz, E., *Information Warfare Principles and Operations*, Boston: Artech House, 1998.

[4] Campen, A. D., D. H. Dearth, and T. T. Goodden (Eds.), *Cyberwar: Security, Strategy and Conflict in the Information Age*, Fairfax, VA: AFCEA International Press, May 1996.

[5] Campen, A. D. (Ed.), *The First Information War*, Fairfax, VA: AFCEA International Press, October 1992.

[6] Von Clausewitz, C., *On War*, London, UK: Penguin Books, Original 1832, 1968.

[7] Schleher, D. C., *Electronic Warfare in the Information Age*, Boston: Artech House, 1999, p. 5.

[8] Scales, R. H., *Certain Victory: The U.S. Army in the Gulf War*, U.S. Army Command and General Staff College, Fort Leavenworth, KS, 1994, pp. 145–147.

[9] Schleher, D. C., *Electronic Warfare in the Information Age*, Boston: Artech House, 1999, p. 2.

[10] Joint Publication 3-51, Joint Doctrine for Electronic Warfare, April 7, 2000.

[11] Johnson, S. E. and M. C. Libicki, Eds. "Dominant Battlespace Knowledge," The Center for Advanced Concepts and Technology, National Defense University, Washington, DC, October 1995.

[12] Krishnamurthy, V., "Emission Management for Low Probability Intercept Sensors in Network Centric Warfare," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 41, No. 1, January 2005, pp. 133–152.

[13] "Information Assurance How Do We Win the War to Protect Information, *The Journal of Electronic Defense*, March 2001, pp. 51–52.

[14] Nicholson, D. L., *Spread Spectrum Signal Design LPE and AJ Systems*, Rockville, MD: Computer Science Press, 1988.

[15] Simon, M. K., et al., *Spread Spectrum Handbook, Revised Edition*, New York: McGraw-Hill, 1994.

[16] Poisel, R. A., and J. L. Hogler, "Global Communications 2010: Military IEW Challenges," Technical Report No. IEWD-RT-930001 (U), USA CECOM RDEC IEWD, Vint Hill Farms Station, Warrenton, VA, February 1992.

Chapter 2

Electronic Support

2.1 Introduction

For the purposes herein, ES is the new name for what used to be called *electronic support measures* (ESM). ES is comprised of those actions to search for, intercept, identify, and locate intentional and unintentional radiation [1, 2]. There are two fundamental purposes to which the information gleaned from intercepted communication signals is applied. Which purpose is determined by the use made of this information, which in turn is determined by the amount of time taken to extract information. If the signals are analyzed over an extended period of time, then intelligence is generated. If the information is put to immediate use, usually not requiring extensive analysis to put it into context, it is called combat information, not intelligence. Of course, after the signal contents have been analyzed, that information can also be used for combat purposes and combat information can be used to generate intelligence. The key difference here is the time it takes to extract useful information. Combat information is used immediately. Therefore, there is little to no time to perform analysis.

ES generates combat information, whereas SIGINT generates intelligence. Reaction to combat information is immediate, whereas intelligence is used to formulate the long-range picture, and thus can take longer to generate.

2.2 Intercept

The intercept of communication signals is one of the principal functions of an ES system. It is important to note that these communication signals are noncooperative—that is, they generally do not want to be intercepted. This is as opposed to the nodes in a communication system, which generally do want to communicate with each other. Noncooperative intercept is a much more difficult problem, especially with digital communication signals [3–5].

2.2.1 Internals Versus Externals

Sometimes information can be extracted from analysis of the *externals* of a signal. Such externals might consist of such factors as the baud rate, RF bandwidth, and modulation type. Determination of such parameters can yield information about the EOB (discussed in Section 2.3) of an adversary, such as the type of military unit one is facing.

The *internals* of a communication signal refer to the information content provided in what was said, or what information was otherwise exchanged (as in modem signals, for example). Meaning can sometimes be extracted from such internals.

2.2.2 Propagation Loss

Radio signals suffer losses as the distance between the transmitter and receiver increases as discussed in Chapter 4 [6]. To the first order, this is due simply to the limited power spreading out in all directions from the transmit antenna. The power density decreases because the surface of the sphere through which the power passes gets larger; thus, the power density must decrease. This is called the attenuation of free space.

Other effects cause further attenuation. Close to the Earth, vegetation and surface irregularities add to the attenuation of free space. Mountains can preclude signal propagation at all, or, in some cases, they can contribute due to a phenomenon called *knife-edge diffraction*. Generally, the higher the elevation of an antenna, the better it can both transmit and receive radio signals, largely because higher antennas tend to be freer of such limitations.

Because signals get weaker with distance between the transmitter and receiver, ES systems generally must be very sensitive. Friendly communication signals typically are closer to the ES system than the targets of interest. The signals from these transmitters are considered interference to the ES system, and they are typically much stronger. The frequency spectrum is channelized in that different parts of the frequency spectrum are divided according to the needs of that spectrum and the technological ability to provide the channels. In the low VHF between 30 MHz and 88 MHz, the military channels are 25 kHz wide. Transmitters designed to operate in this range must have the majority of their power concentrated in that 25 kHz. This is not all their power. Theoretically it is impossible to have all the energy in a signal within a limited bandwidth while simultaneously having a definite start and stop time for the transmission. Therefore, there is energy from those transmitters that is outside of the 25 kHz channels. In fact, significant energy can extend to several megahertz on both sides of the channel. This energy further exacerbates the interference problem.

Therefore, the ES system must be able to intercept weak signals even though there are very strong signals nearby causing interference and those signals are not necessarily close to the frequency that the ES system is trying to intercept. The parameter that indicates how well an ES system can operate with this near-far problem is its dynamic range. It is difficult to achieve adequate dynamic ranges for ground-based ES systems. The *required* dynamic range might be 120 dB, whereas *achieving* 70 dB can be difficult. Airborne ES systems receive the weak target signals much better than ground systems, so their dynamic range requirements are less, although they are still challenging. In fact, since the airborne systems are higher, they see more of the friendly interfering signals than the ground systems do, so their interference problem can be worse.

2.3 Geolocation

The location of targets is usually considered critical combat information. These locations can be used to track target movement over time as well as to indicate groupings of types of targets into a particular region. Such groupings are often indicative of unit type, intentions, and plans. The amassing of units in a close area, detected and located by EW means, might indicate a specified type of activity [7, 8].

Geolocating communication emitters, known as *position fixing* (PF), is one of the more important functions of ES systems. PF is the process of locating in three-dimensional space the origination of the radiating entity. Such information has several uses. Depending on the use to which the resultant fixes are applied, it determines to a large extent the accuracy requirements. Some of the requirements are much more stringent than others, depending on how precisely the resultant location is required.

By determining the location of emitters in the battlespace, it is frequently possible, normally by combining this information with other information, to determine the EOB. This information is an indication of who, what, when, and where about the opposing force. It also attempts to determine the intentions of the opposing force in terms of their mission. Although knowing the location of entities accurately is useful, it is frequently not required to know their location as precisely as some other applications. Finding their location for EOB determination on the order of 1 km is generally adequate for this application.

Sometimes weapon platforms have unique communication signals associated with them to exchange command and control information. Locating and identifying these signals facilitate identification of the weapon system. Finding these weapon platforms is useful for focusing the direction of jamming signals as well as, as typified by the *surface-to-air missile* (SAM) case, for threat avoidance.

In these applications, precise locations are not necessary either with accuracy requirements on the order of 1–5 km being adequate.

If emitters can be located accurately enough, then that information can form the basis of providing targeting information for attacks by indirect weapons (artillery and missiles mostly). For such dumb weapons, accurate targeting is required—often less than 100m. The location of emitters, however, does not necessarily mean the location of the weapon platform. It is possible to position the transmitting antenna of an emitter quite a distance from the platform. In this case, the resultant geolocation must be accurate enough so that the indirect weapons can hit the target with a reasonable number of dumb shells. Of course, for smart weaponry, such precise locations are not necessary, as the round itself finds the precise location once it gets close enough. Thus, for smart weapons, less precise geolocations are adequate—on the order of 5 km.

In *operations other than war* (OOTW), locating communication emitters might form the basis for tracking individuals or units. Command and control of mobile forces almost always require the use of mobile communications. These communications might take advantage of the commercial infrastructure, such as cellular phone systems. They might be facilitated by *citizen band* (CB) commercial transceivers. They may also be accomplished with transceivers designed specifically for that purpose. Whatever the means, sometimes it is useful to geolocate transmitters even in OOTW.

In the United States the *Federal Communications Commission* (FCC) uses the techniques discussed herein to find illegally operating transmitting stations. To operate an RF transmitter above some power threshold in all but a few frequency bands [called the *instrumentation, scientific, and measurement* (ISM) bands] in the United States, a license is required. This license is issued by the FCC. Operating without a license is illegal, so one of the tasks of the FCC is to geolocate offending transmitting sites.

Another commercial use of geolocation is the cellular phone system and 911 emergency calls. If a 911 call is made from a cellular handset, the signal can frequently be detected at more than one base station. These base stations can triangulate to locate where the emergency call is coming from, so the emergency vehicles will know where to go without relying on the caller to provide that information.

2.4 Triangulation with Multiple Bearings

One of the most common ways to accomplish emitter geolocation is by *triangulation* [9]. In this, the direction of arrival of an incoming wave front, called the *line of bearing* (LOB) or just *bearing*, also LOP, is determined at two or more

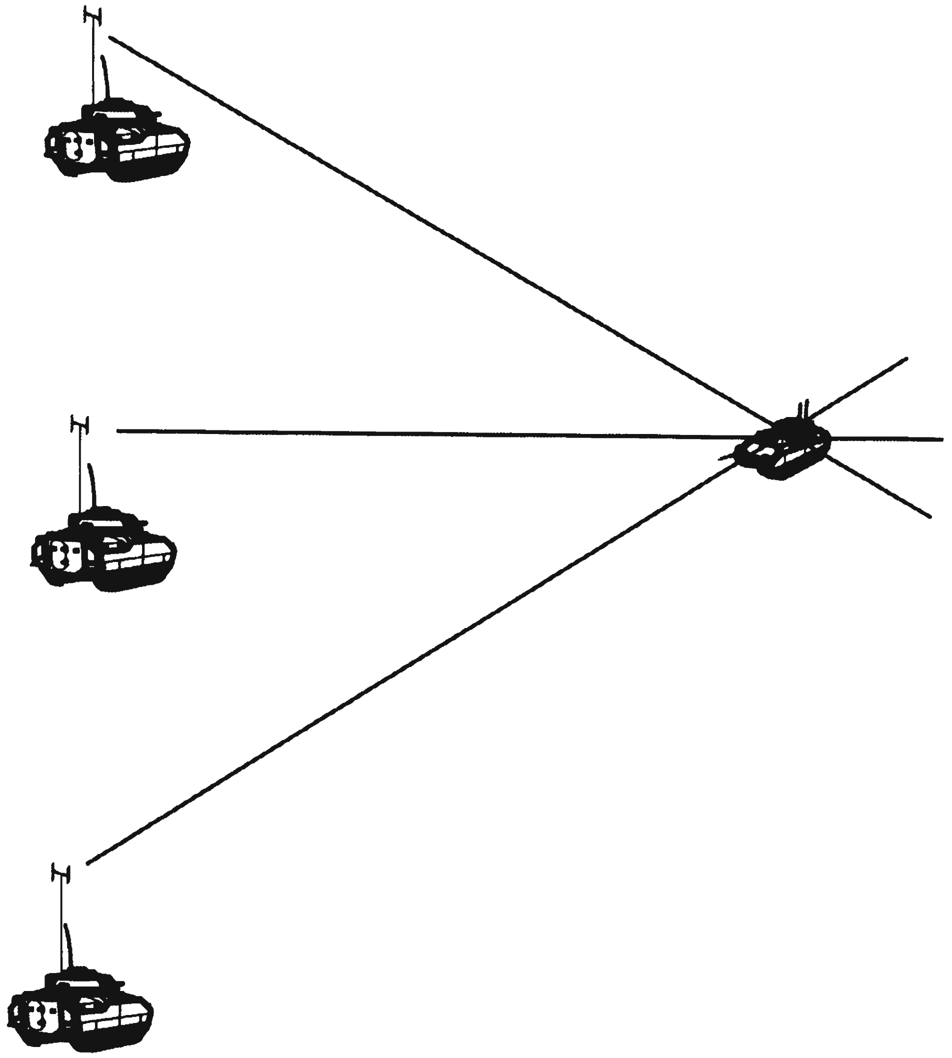


Figure 2.1 Example situation where direction finding is used.

sites. Where these bearings intersect, which they will always do unless they are parallel lines, is the estimated location of the emitter. A typical example is shown in Figure 2.1, where three sites are obtaining bearings on the radio in the tank shown.

There are always errors present in the measurement of the bearings. Such errors can be systematic errors, errors caused by violation of the assumptions about the wave front, or noise induced. The principal assumption is that the wave front is planar—that is, it arrives at the receiving system as a plane wave. At sufficiently large distances between the transmitter and receiver, this assumption is usually valid relative to the curvature of the wave caused by its spherical propagation. What is frequently not true, however, is that the wave is planar because of unexpected perturbations in the path that the wave takes. Multipath, for example, can cause more than one wave to arrive at the receiving site, probably with different amplitudes, phases, and direction of arrival. These errors cause uncertainty in the calculation of the location of an emitter. Because of this a measure of confidence is usually computed along with the geolocation, or "fix." This measure of confidence is usually a contour describing a region within which the emitter is located with a prescribed measure of probability. If the region is a circle, then it is called the *circular error probable* (CEP). The CEP is the radius of the circle. If it is an ellipse, then it is called the *elliptical error probable* (EEP). The semimajor and semiminor axes as well as the tilt angle of the major axis describe the EEP. The most common probabilities employed are 50% and 90%.

Typical operational accuracy for RF direction-finding equipment depends on the frequency range of interest because of the environmental effects on signal propagation. The HF frequency range, for example, can exhibit considerable reflections from and refraction through the ionosphere and terrain. The *instrumental accuracy* refers to the accuracy of the system itself, without the effects of environmental factors while the *operational accuracy* is a measure of the expected performance once the equipment is put into operational use. There can be considerable differences between the two, and complicating the matter is that the operational accuracy is usually determined by test in field conditions. Unless the field conditions are varied over all of the expected operational environments, there will be biases in the measurements. For modern equipment in the VHF and UHF range, the field accuracy is typically 4° to 5° *root mean square* (RMS). In the HF range, typical operational accuracy ranges from 10° to 25° RMS.

Whereas putting PF assets on airborne platforms normally has significant benefits, putting such assets on UASs has several additional advantages. The first of these is the required accuracy. Referring to Figure 2.2, if the PF hardware is

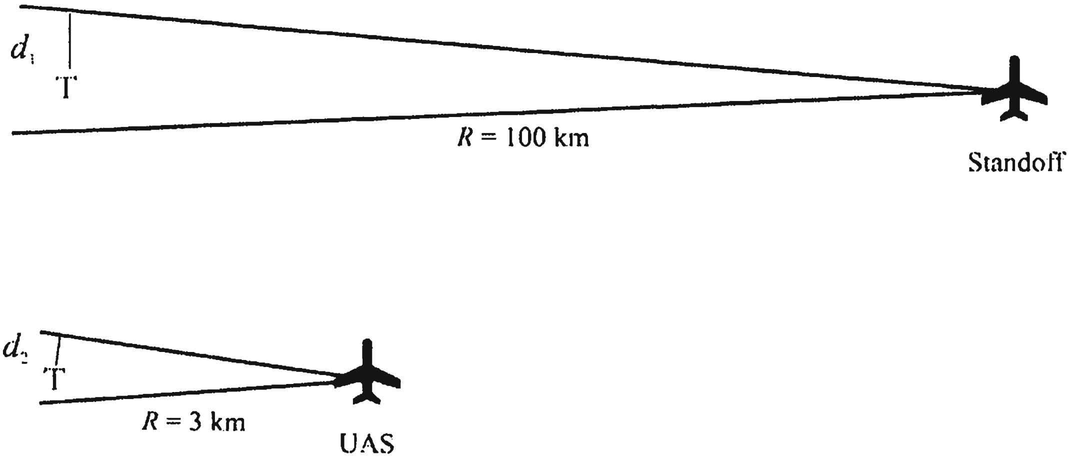


Figure 2.2 Comparison of the miss distance of a 2° accuracy PF system from a standoff platform versus a penetrating UAS. T is the location of the target.

mounted on an airborne standoff platform that is at a range of, say, 100 km from the target of interest, the situation would be as shown at the top of Figure 2.2. Assuming the (in)accuracy of the PF system is 2°, then d_1 = 100 sin 2° = 2.5 km. On the other hand, if the PF system were on a UAS as shown at the bottom of the figure, it can overfly the adversary's air space, getting much closer to the target. When the range to the target is 3 km, then the same PF system at 2° accuracy produces d_2 = 105m.

An additional benefit of overflying the target area is the proximity of the UAS to the target and the requirement to obtain multiple LOPs to compute a fix. With more than one system available to obtain simultaneous LOPs on a target, then a single opportunity is adequate to compute a fix. If there is only one asset available, it is necessary for the platform to move some distance to obtain a second (or more) LOP on the target. If the platform is standing off 100 km or so, it will take some time for the system to move an adequate distance to obtain the second LOP. For a good fix, the system would have to move on the order of 100 km before the second LOP could be taken. At 100 km/hr velocity this would take an hour. Chances are good that the target will not be emitting an hour later. On the other hand, if the UAS is close to the target, it would have to move only 3 km for the second LOP. At 100 km/hr this would take only a minute and a half or so.

Alternatively, less accurate PF hardware is required to achieve comparable target location accuracy. If the requirement is for a miss distance of no more than 250m, then a standoff system at a range of 100 km would need PF accuracy at 0.14°, while a penetrating UAS system could achieve this miss distance at a range of 3 km with a PF accuracy of a generous 4.8°.

The second advantage is that the system operators can be remotely located on the ground in a relatively safe location. Operations could continue even though the supported force is moving because the UAS can move ahead of them or with them. There is no need for on-the-move operation of the ground segment. If two ground segments were available, they could leapfrog in their operation keeping continuous coverage of the area.

A third advantage is that the sensitivity of the hardware to achieve this accuracy is substantially less. The targets are much closer so for a given amount of emitted power, more power is received by the UAS system than the equivalent standoff configuration. This advantage, however, in some cases will be a disadvantage. Lower sensitivity means smaller instantaneous coverage regions. The UAS may need to move from one area to another to cover the entire region of interest.

One of the more significant shortcomings of overflying targets in a UAS is the depression angle. Standoff systems need to be calibrated in azimuth only, limiting the depression angle requirements to the horizontal plane. A UAS at 3,000m altitude *above ground level* (AGL) in Figure 2.2 has a depression angle of about

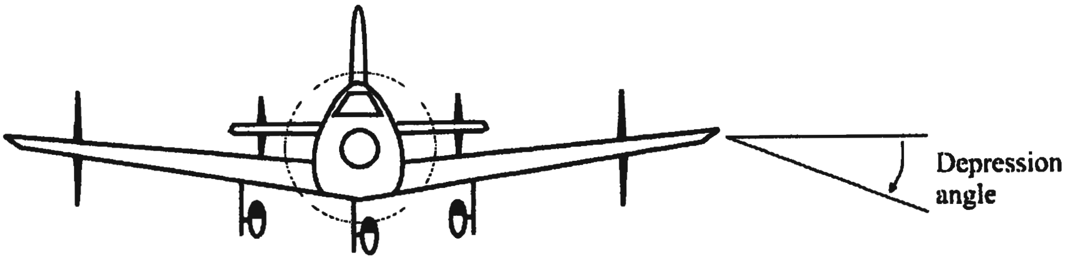


Figure 2.3 Depression angle associated with airborne systems.

45°, so substantially more system calibration is required. In addition, the dipoles or monopoles usually used for airborne applications have limited response beneath the aircraft, thus decreasing system sensitivity. The depression angle is shown in Figure 2.3.

Lastly, putting ES assets among the target array minimizes the difficulty of looking through the friendly array of emitters. With U.S. forces using SINCGARS radios, which hop around the low VHF spectrum, considerable friendly RF fratricide occurs. Standoff ES systems must look through this crowded spectrum to find targets.

All of these advantages, among other things, lower the cost of ownership of the system by decreasing the acquisition cost, making a simpler system and thereby increasing the reliability and decreasing the amount and number of spare components. Also decreased are the maintenance personnel costs, since fewer will be needed.

2.5 Deployment Considerations

Ground-based systems are most useful when extended ranges are not required. Such situations consist of protecting early-entry forces by monitoring the region immediately surrounding a landing area. If ground-based systems must be used to cover terrain at extended ranges, then they should be located on the highest terrain possible. These locations must provide for access for such items as fuel, resupply, and rations even under adverse weather conditions. Reverse slopes should be considered with the antenna just reaching over the top of a ridge or hill. Care should be used in such cases, however, because geolocation accuracy can be significantly affected by close proximity to the ground [10–12].

Since antenna height is important in such applications, historically ground-based systems have had to stop and erect antennas in order to get much range. Alternatively, receiving equipment mounted on UASs and interconnected to ground systems in real time can provide extended coverage while still providing protection for the system operators on the ground. On-the-move operation is even possible in this configuration.

It is possible to remote ES receiving front ends from the back ends, where the operators are located. The interconnects could be terrestrial line-of-sight data links, satellite links, or relayed through a UAS configured for the task. Such configurations not only are safer for the system operators but they facilitate some types of operations that are not possible otherwise. If there is language translation necessary to support an operation, the linguists could all be located at the same location, thereby easing the language mix and skill level issues.

2.6 Electronic Mapping

If adequate geolocation accuracy can be achieved and adequate emitter identification can be accomplished, then it is possible to develop and maintain a map of the electronic targets. The difficulty is in obtaining the adequate accuracy. Such a map could indicate the up-to-date disposition of adversary forces as well as possibly show interconnected networks. As will be discussed in Chapter 7, identification comes in many forms, and such identification would add valuable information to the map.

Such a display might be a map or terrain overlay indicating the locations of targets calculated from data obtained from the PF system. If the data is available, in addition to the locations, the type of target could also be indicated. The type could be anything that is available from the ES system, such as frequency range, modulation type, or enemy or friendly. The type of unit could be indicated if associations have been made from several intercepts from the same location of multiple types of emitters. Groupings of targets within particular regions might indicate that some sort of specific activity is happening or about to happen.

Groupings of emitters within a region could also be used as a tool for collection management. Geographical filtering could be invoked so that the system resources are only applied to intercepts on targets emitting from that region. Also, other sensors could be deployed to that region to confirm information and otherwise collect information.

2.7 Common Operational Picture

ES can support generation of the *common operational picture* (COP), not only by geolocating targets, but also, in some cases, identifying those targets by unit type, or in some cases, by identifying the specific unit. The latter can only occur if specific emitter identification is accurate enough, however. If a target were intercepted previously and the technical capability is present to identify the target as one that was intercepted previously, then the system can automatically indicate that the same target is present. If the target is in a new location, then tracking the target is possible, and this too can contribute to the COP by indicating such movement.

Detection of particular types of emitting targets within a geographical region can sometimes indicate intentions, movements, or other significant battlespace activities. Certain weapon platforms sometimes have unique combinations of emitters that can be intercepted and geolocated indicating that a specific weapon platform is present. It also can sometimes indicate that the weapon platform is in a particular state, for example, preparing to fire artillery. This also contributes to the COP generation and maintenance function.

Until reliable automated translation of natural language is available, a person will be required to translate communication intercepts. That is not the situation with computer-to-computer digital signals. Since these signals are processed at both ends by machines anyway, it makes it possible to potentially automatically process them in an ES system.

2.8 Operational Integration with Other Disciplines

One of the more significant capabilities provided by communication ES systems is their ability to cue other intelligence systems. An EO/IR sensor has a very narrow field of view, at least relative to an ES system. These sensors, however, can be cued by an ES system with a geolocation capability, which reduces the search range required of the EO/IR system by two to three orders of magnitude or more (depending on implementation). An EO/IR system on a UAS has a total scan width of about 10 km or so on the surface of the Earth, while the instantaneous imaging capability is about 0.5 km or so. An ES system on that same UAS has a substantially larger field of view, which can be used to focus the EO/IR payload quickly to targets of interest. TOCs and other formations of interest are frequently guarded by tactical units, which must communicate in order to function. These communications are the targets of UAS ES systems, and they can be used to efficiently cue the EO/IR sensors [13–15].

Radars, such as *synthetic aperture radars* (SARs) or *moving target indicating* (MTI) radars, are sensors with a broad coverage area. They cannot, however, tell much about what an entity is beyond basic indications, such as tracked vehicle versus wheeled, for example. ES systems can sometimes provide that information. Combining this information with that from a radar can indicate that a particular unit is moving or has moved to a new location or that it has been combined with other units. This type of information is useful for event detection, unit tracking, and the like. Like other systems that provide geolocation information, ES systems can provide cues for HUMINT collection by indicating the location of particular units, and, sometimes more importantly, where adversary units *are not*.

There is also the need for sensor registration since multiple sensors are typically not collocated. Registration removes (to the extent possible) errors in locating targets due to misalignment of the sensors being used to provide the location information.

2.9 Support to Targeting

Another use for ES, as with other intelligence disciplines, is to support targeting. This is the process of identifying and verifying targets as being high priority—high enough to attack with (usually) indirect-fire weapons, including EA assets. ES sensors are considered wide-area sensors, as opposed to, for example, TV cameras that only cover a small area at a time. As such, ES sensors are most useful for targeting purposes by cueing other confirming assets such as EO/IR sensors or humans [16].

2.10 Concluding Remarks

An overview of the operating principles of communication ES systems was presented in this chapter. ES systems provide two fundamental capabilities to military commanders. The first is the generation of combat information. This information is of immediate battlefield use and requires little or no analysis to be useful. The other type of information is intelligence. The systems to collect these types of information are not necessarily different. It is the utilization of the resultant data that defines the type of information, not the collection means. The amount of data collected, however, may vary depending on the tasks at hand. The ES system design principles described herein apply in both situations.

References

[1] Waltz, E., *Information Warfare Principles and Operations*, Norwood, MA: Artech House, 1998, p. 214.

[2] Schleher, D. C., *Electronic Warfare in the Information Age*, Norwood, MA: Artech House, 1999, pp. 1–2.

[3] Torrieri, D. J., *Principles of Secure Communication Systems*, 2nd Ed., Norwood, MA: Artech House, 1992, pp. 291–362.

[4] Wiley, R. G., *Electronic Intelligence: The Interception of Radar Signals*, Dedham, MA: Artech House, 1982.

[5] Neri, F., *Introduction to Electronic Defense Systems*, Norwood, MA: Artech House, 1991, pp. 30–32.

[6] Stark, W., et al., "Coding and Modulation for Wireless Communications with Application to Small Unit Operations," accessed March 2008, http://www/eecs.umich.edu/systems/TechReportList.html.

[7] Torrieri, D. J., *Principles of Secure Communication Systems*, 2nd Ed., Norwood, MA: Artech House, 1992, pp. 335–366.

[8] Wiley, R. G., *Electronic Intelligence: The Interception of Radar Signals*, Dedham, MA: Artech House, 1985, pp. 107–134.

[9] Torrieri, D. J., *Principles of Secure Communication Systems*, 2nd Ed., Norwood, MA: Artech House, 1992, p. 364.

[10] Torrieri, D. J., "Statistical Theory of Passive Location Systems," *IEEE Transactions on Aerospace Electronic Systems*, Vol. AES-20, No. 183, March 1984.

[11] Adamy, D., *EW 101: A First Course in Electronic Warfare*, Norwood, MA: Artech House, 2001 pp. 145–147.

[12] Neri, F., *Introduction to Electronic Defense Systems*, Norwood, MA: Artech House, 1991, pp. 271–332.

[13] Hall, D. L., *Mathematical Techniques in Multisensor Data Fusion*, Norwood, MA: Artech House, 1992.

[14] Waltz, E., and J. Linas, *Multisensor Data Fusion*, Norwood, MA: Artech House, 1990.

[15] Neri, F., *Introduction to Electronic Defense Systems*, Norwood, MA: Artech House, 1991, pp. 457–472.

[16] Waltz, E., and J. Linas, *Multisensor Data Fusion*, Norwood, MA: Artech House, 1990, pp. 110–113.

Chapter 3

Electronic Attack

3.1 Introduction

There are two fundamental types of information denial. We can deny an adversary our own information and we can deny that adversary his or her own information. The former is called *information protection* and the latter is called *information attack*. There are a variety of ways to execute information attack and protection, such as camouflage, concealment and deception, and EW. The technologies described in this chapter are limited to this last category, attacking an adversary's communication systems to deny information transport. These principles are readily extended to protecting one's own communication from exploitation by an adversary's communication exploitation systems, so this case is covered as well. The other forms of information denial are not covered herein.

3.2 Communication Jamming

The fundamental task to be accomplished with jamming was illustrated with several examples in Chapter 1. Simply put, a communication jammer tries to deny communication over RF links. A jammer attempts to accomplish this by putting unwanted signal energy into the receivers in the communication system. This unwanted energy, if strong enough, will cause the receivers to demodulate the signal from the jammer as opposed to the communication transmitter. Assuming that the jammer signal is not a replica of what was transmitted, communication is denied on the RF link [1–4].

Communication jammers can be used to deny information in three fundamental ways: (1) denying an adversary the ability to talk to another element, thereby limiting command and control, (2) jamming an ES/SIGINT system, also known as *communication screening*, and (3) deceiving an adversary's ES systems.

Jamming radio signals has been around almost as long as radio signals themselves. There are several ways one can perform EA against an adversary's RF communication systems. Jammers come in a variety of configurations as well. A jammer that operates from within the friendly held battlespace is called a *standoff* jammer. One that operates within an adversary's held battlespace is called a *stand-in* jammer. A jammer that attacks the target's carrier frequency only is called a *narrowband* jammer, while one that emits a broad range of frequencies simultaneously is called a *barrage* jammer. Due to fratricide considerations, barrage jammers are rarely used in standoff configurations.

Jamming communication nets is always part of a larger military or political maneuver. It is integrated with the battle planning, just as other indirect fire weapons are integrated. The use of jamming must be synchronized with other battlespace activities to maximize its utility and impact on the outcome. The two primary intents of communication jamming are to disrupt an adversary's command and control process—thus, it attacks the connection between the decision makers and the sensors and/or the implementers of attack. It attacks the neck of an adversary, which carries bidirectional information to and from the decision makers.

Jamming commercial TV and radio broadcast stations may be included in an EA campaign. TV stations are frequently used to broadcast propaganda to noncombatants in an area, thereby attempting to influence the local populace against a friendly operation. Jamming such signals precludes this propaganda from getting through. Radio Free Europe was jammed for many years by the Soviet Union as are current attempts to communicate with the Cuban populace by the United States.

One of the goals of communications EA is to interrupt communications, or otherwise preclude effective communications, between two or more communication nodes. While digital communications are overtaking the analog forms, the latter is still expected to be around for quite some time, so effective EA against such systems is expected to continue to be required. Also, many of the same principles apply for both forms. The ability to interrupt analog speech signals depends on several factors. Shown in Figure 3.1 are curves for several degrees of interruption along with the rate of that interruption [5]. The intelligibility of the resultant speech is the dependent variable here. The parameter of the curves is the amount (percentage) of the interruption, varying between 25% and 87.5%. The way to interpret this chart is illustrated in Figure 3.2. Suppose there is a 10-second message. The case of the rate = 1 and 25% interruption is shown in Figure 3.2(a). On the other hand, the case for rate = 10 and 25% interruption is shown in Figure 3.2(b). Other rates and interruption fractions are interpreted similarly. It is interesting to note that while the total interruption time shown in Figure 3.2 is the same, the effect on the intelligibility is different by about 20%. Grouping the interference into the same location in time, as it is in Figure 3.2(a), has a greater

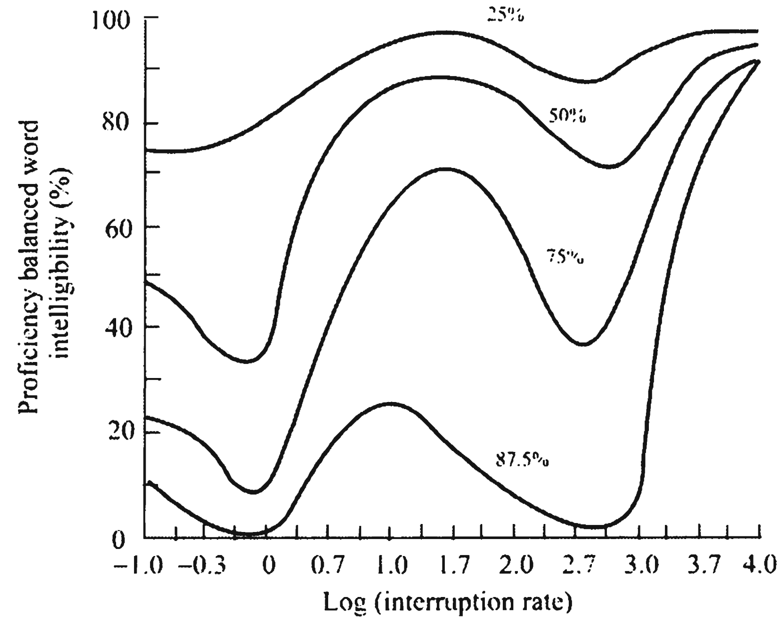


Figure 3.1 Intelligibility of interrupted analog communications as a function of the rate of interruption (expressed as the logarithm) and the degree (percentage) of the interruption per event. (*Source:* [5]. © IEEE 1992. Reprinted with permission.)

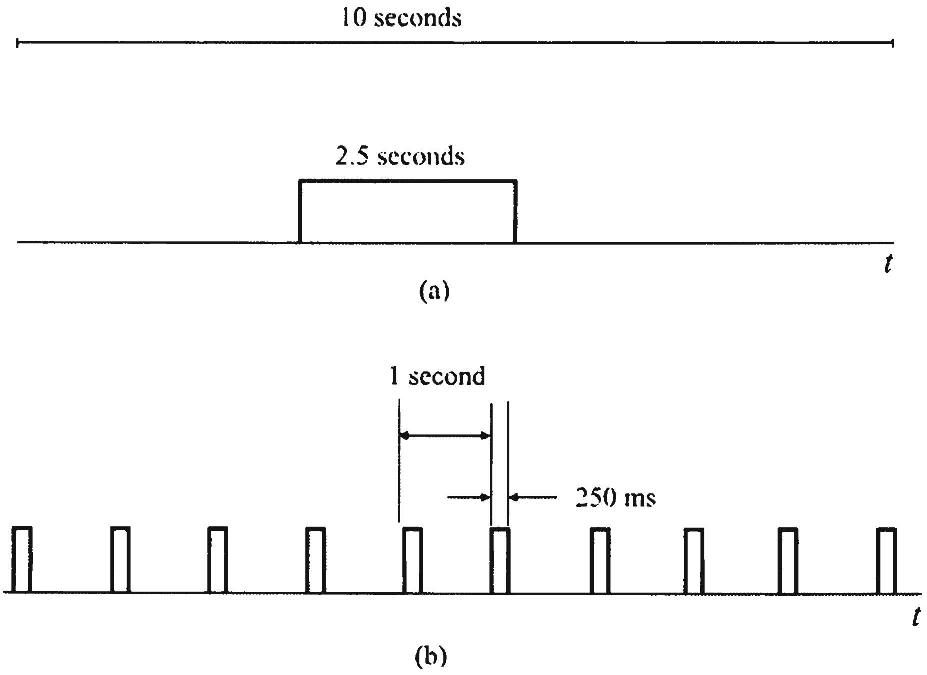


Figure 3.2 Interpretation of the interruption rate and amount terms for Figure 3.1: (a) rate = 1 at 25% interruption, and (b) rate = 10 at 25% interruption.

effect than spreading it out as in Figure 3.2(b). Furthermore, this effect is not the same for all combinations of rate and degree, although the general shape of the characteristics is consistent, so design tradeoffs must be carefully considered to optimize the effects of a jammer.

Experimentation has shown that a percentage of interruption of approximately 30% causes significant degradation of analog voice communications. Therefore, articulations above about 70% are required in order for tactical communications to be effective. It also indicates that a single jammer can simultaneously handle about three analog targets, switching the jamming signal quickly between the three signals. The particular effects are subject to the repetition rate of the interruptions due to the jammer as shown in Figure 3.1. Rates of adequate length (say, 25%) on the order of one interruption per transmission are adequate to reduce the articulation rate sufficiently to preclude acceptable communication.

International treaties may preclude jamming international satellites for obvious reasons. Communication satellites are carrying more international communication traffic; the *international telecommunications satellites* (INTELSATs) carry thousands of telephone calls per day all around the world. Jamming one of the transponders on one of these satellites could take out telephone calls from several countries at the same time, some of which would probably be friendly. In fact, that friendly communication may very well be diplomatic or military communications attempting to forward the friendly cause. Thus, extreme care must be taken before attempting to negate such communication paths.

The amount of jamming signal power at the input to the receiver to be jammed relative to the signal power at that input determines the effectiveness of the jamming scenario. The amount of this jamming power depends on several factors:

1. The *effective radiated power* (ERP) of the jammer;
2. The ERP of the communication transmitter;
3. The orientation of the receiver antenna relative to that of the jammer antenna;
4. The orientation of the jammer antenna relative to that of the receiver antenna;
5. The intervening terrain.

If the antenna at the receiver has directionality, then it probably would be pointed more or less in the direction of the transmitter to maximize received energy. Unless the jammer were within this beam, it is at a disadvantage because the receiver antenna has less gain in the direction of the jammer. The jammer power, therefore, must be increased to overcome this disadvantage. The actual parameter of importance is the *J/S ratio* (JSR), where *J* is the average jammer power and S is the average signal power; both of these power levels are measured at the receiver. We will denote the JSR here by ξ. This ratio is most often given in decibels:

$$\xi_{\text{dB}} = 10\log(\xi) \tag{3.1}$$

3.3 Jammer Deployment

A jammer can be expendable. In that case, it might be delivered to its operating site by a soldier who emplaces it by hand, or it might be delivered by some other means such as deployed from a UAS or shot into place out of an artillery cannon or rocket. Whatever the delivery mechanism, expendable jammers obviously must be economical so that they can be lost. They also must be controllable. Since tactical campaign plans almost always change once the operation is under way, one does not want to emplace active jammers in an area, and then expect one's own soldiers to communicate in that area if they must pass through there with the jammers active [6, 7]. We discuss distributed jammer configurations and their performance in Chapters 14 and 15.

An array of expendable jammers can be formed by any of the emplacement methods mentioned, air delivered by UASs, hand-emplaced, or artillery-delivered. The array defined by a single artillery shell would be more or less a straight line, whereas an arbitrary pattern can be constructed by the other two methods. The jammers would be placed nominally 500m apart. An area seeded with such jammers could have arbitrary width and depth. Placed within a valley, for example, as an adversary force moves through the valley, they could be precluded from communication by radio among themselves as well as with higher echelon command elements.

Deploying communication jammers onboard penetrating UASs has several advantages. The first of these is fratricide avoidance. Standoff jammers, being closer to friendly radios than targets, can inject unwanted, interfering energy into friendly communications. By placing the jammer in the adversary's battlespace, nominally all communications can be viewed as targets so the fratricide problem disappears. This does not say, however, that all such communications should be jammed—just that friendly fratricide can be minimized this way.

A second advantage of UAS jammers is their minimal vulnerability. One of the problems with manned jammers is their survivability once they are put into operation. Such high-power transmitters are relatively easy to find both in frequency and geolocation. This makes them particularly vulnerable to enemy attack, from direct fire as well as indirect fire weapons, either of which could involve homing weapons. Jammers mounted on UASs are not stationary. In order to eliminate such a system, it would need to be tracked. Furthermore, there are no operators physically collocated with the jammer. These aspects make a UAS-borne jammer substantially less vulnerable and more survivable (to include the operational personnel).

A third advantage of UAS jammers is the relatively smaller amount of effective radiated power required to accomplish the jamming mission. Being substantially closer to the targets than their standoff counterparts, less power from the jammer is required for the same jamming effectiveness. Among other advantages, this makes system design easier, which leads to higher reliability and improved operational utility.

In addition to possibly being closer to the jamming targets, the propagation conditions are better—for close-to-ground signals the signal level falls proportionally to R^{-4}, whereas an air-to-ground link falls proportionally to R^{-2} or R^{-3}. These equate to losses of 12 dB, 6 dB, and 9 dB per octave distance, respectively. On the other hand, electronic fratricide is possibly increased when using a UAS for jamming.

UAS-mounted jammers can be controlled by ground operators substantially distant from the UAS platform. These operators could actually be collocated with the fire support officer in the battalion or brigade TOC where real-time coordination of EA with other forms of indirect fire can occur. A TOC is where tactical battle management occurs.

Another form of UAS jammer could take advantage of micro-UASs. Placing dumb, barrage jammers onboard very small UASs and launching them by hand over a small region could prevent an adversary from communicating for, say, 5 minutes, while a platoon-sized operation moves through that region.

3.4 Look-Through

Look-through is incorporated in jammers to maximize the utility of the available power from them. Once a narrowband target is being jammed, it is normal for that target to change to a prearranged backup frequency that is not being jammed at the moment. The jammer must therefore monitor the target transmission to tell if it is still trying to communicate at the first frequency. A receiver collocated with the jammer cannot monitor at the frequency that is being jammed simultaneously. If it did, it would either be destroyed by the high-power signal at its sensitive front end, or at least it would be desensitized enough so that it could not hear the target. Therefore, for a short period the jammer is turned off so that the monitor receiver can measure the target energy at the frequency. This is called *look-through*. Typically several tens of milliseconds are adequate to determine if the target is still transmitting so the jamming signal need only be disabled for this amount of time.

If look-through were not incorporated, the jammer would waste a good deal of time jamming a target that was no longer trying to transmit at that frequency. This degrades the performance of the jammer in the number of signals it can jam per mission. Techniques incorporated to maximize the utility of the jammer, such as look-through, are called *power management.*

3.5 Analog Communications

The amount of jammer power required to prevent communication depends on the type of modulation used. Some types are more easily jammed than others. Analog voice, either AM or FM, for example, requires significantly more jamming power than most digital communications—unless the latter contains considerable FEC coding [8, 9].

Modulating a high-power carrier signal with an FM noise signal is the old standby for jamming analog communications. It has been shown several times that this is the best waveform to use against FM targets. Such targets are very plentiful in the VHF band. It has long been known that analog FM communications exhibit a threshold effect wherein if two signals are impinging on a receiver at the same time at the same frequency, the stronger one will dominate the receiver to the more or less complete rejection of the weaker one. The same will happen if one of the signals happens to be a jammer with FM modulation. It would then be expected that in such systems, a jammer broadcasting an FM signal only need be higher in power by some small amount in order to capture the receiver. This indeed does happen. It also occurs in AM systems, however. Shown in Figure 3.3 is the articulation index plotted versus ξ (noise is considered negligible relative to the jamming signal in this case) [10]. An articulation index of 0.3 or less implies

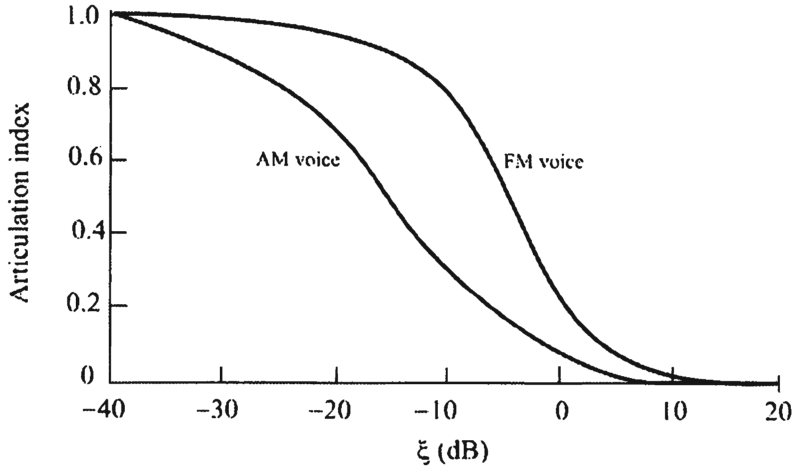


Figure 3.3 Example showing the effects of ξ on intelligibility in analog communications. (*Source:* [5]. © IEEE 1992. Reprinted with permission.)

unacceptable performance, between 0.3 and 0.7 is marginal performance, while above 0.7 is an acceptable performance (all from a communicator's point of view). Note that these curves transition from acceptable articulation to unacceptable fairly sharply, although the capture effect of FM causes it to transition somewhat faster, albeit later. For analog FM, –6 dB JSR or higher is usually considered adequate for disruption of the transmission. At that point, as seen in Figure 3.3, the articulation index is about 50%. The equivalent for AM occurs at about –15 dB, making the latter somewhat less tolerant to jamming than FM.

Other forms of modulated signals are optimal for other forms of modulation, but FM by noise is not a bad choice for most target modulations of interest. Of course, if one knows exactly the waveform that needs to be jammed, one can design an optimal waveform just for that target. This is not usually the case, however.

3.6 Digital Communications

One of the design criteria for digital communication systems is the BER environment in which they are to operate. Theoretically, if the BER is raised above this level, then communication degradation will occur—usually initially in the form of information transport slowdown and, if the BER gets high enough, complete denial of information exchange [11, 12].

A digital signal need not be jammed continuously in order to generate a high BER. A BER of 0.5 can be achieved against a continuously broadcast digital signal (with adequate signal levels) by only jamming 50% of the time. It can thus be concluded that denying information on digital communication signals should be considerably easier than on analog signals.

In addition to traditional jamming techniques against digital (computer-to-computer) communications, it is also possible to load the transport channel with information to the point where it becomes completely saturated. One would conceivably want to insert bogus data over the channel, thus disallowing the intended information to get through, or at least slow the transport of valid information. Most computers are vulnerable to such data overload, since valid information can be difficult to separate from bogus data, as long as the protocol and format are correct. An invalid report that indicates that there has been a tank battalion spotted at coordinates (x, y) is difficult to detect in a system that is expecting to see reports on tank battalions and their locations.

This is one of the downfalls of the Information Age. Unless and until computers can reason with data, as opposed to just processing it, they will be vulnerable. Indeed, humans are vulnerable to information overload, although to a lesser degree than computers.

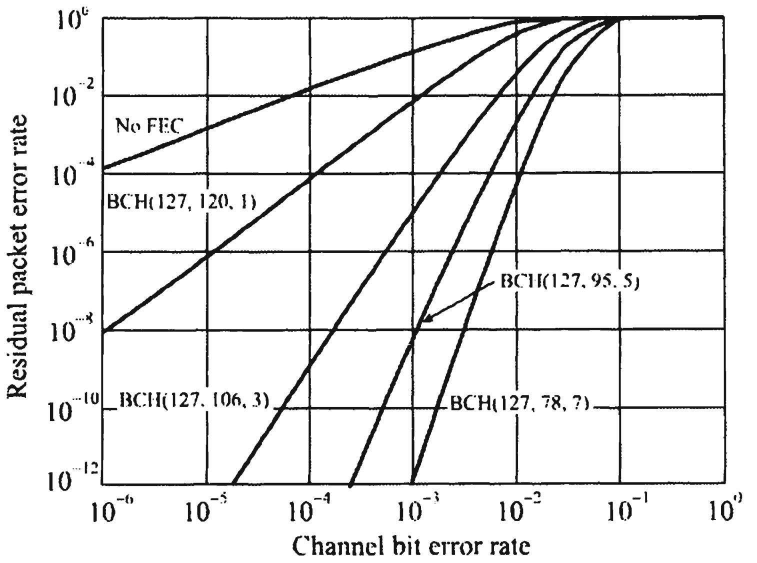


Figure 3.4 Error performance of some BCH codes. Also shown is the case of no FEC.

Channel coding can be applied to digital signals in order to tolerate higher levels of noise, interference, and jamming. Coding, however, applies to cases where the BER is already rather low (say, less than 10^{-3}). At a BER level of 10^{-1}, there is virtually no protection supplied by most simple coding schemes. The unencoded BER is about the same as the channel BER. It is, however, true that if desired the communicator can apply sufficient coding to beat any jamming scheme, but at that point there is so much coding that there is virtually no information throughput.

Bose-Chauduri-Hocuenghem (BCH) codes are popular error correcting, variable length codes that are easily decoded. They are widely used in communication systems. The BER performance for some typical BCH codes are illustrated in Figure 3.4. Note that as the channel BER approaches 10^{-1} or so, the residual packet error rate, that is, the BER after decoding, is also about 10^{-1}, which is too high for reliable communications.

3.7 Narrowband/Partial-Band Jamming

For targets at individual frequencies, a narrowband jammer would be used. The jammer at the frequency used by the targeted communication net transmits a powerful noise signal, or perhaps one or more tones if the target is *binary frequency shift key* (BFSK) (modulations are covered in Chapter 6). The intent is to have the jammer overpower the intended communication signal at the intended receiver. Such a technique is frequently used for standoff jammers to minimize friendly fratricide associated with barrage jammers [13, 14].

Tones or narrowband noise is useful against DSSS targets as well. DSSS communications systems, as exemplified by the IS-95 cellular radio standard, are particularly sensitive to strong signals close to the receiver: the so-called *near-far problem* inherent in DSSS systems. If an adequately powerful narrowband signal can be located close enough to the intended receiver, the processing margin in the receiver can be overcome, causing unreliable effects to occur in that receiver. When the processing gain (defined later) is overcome by the jammer, noise leaks through the demodulation process and raises the noise floor at baseband, thereby decreasing the detected SNR. The SNR is a common parameter used to specify communication system performance. It is used directly for analog communications and determines the BER for digital communications.

It is possible to utilize the same broadband transmitter and antenna to jam more than one target signal at a time. As mentioned earlier, such techniques carry the appellation power sharing. One exciter is required for each such signal since the exciter determines the frequency to be jammed. The power per signal jammed in this scenario theoretically decreases approximately as the square of the number of signals being jammed with the common power amplifier, however. In practice the power per signal decreases faster than the square of the number of signals because of mismatch issues at the output of the power amplifier. In addition, this technique does not work if the output of the power amplifier is tuned to the jamming frequency unless there is a separate tuned filter at the output of the amplifier for each jamming signal.

It is also possible to time-share a jammer to jam more than one signal essentially simultaneously. Such techniques are called *time sharing*. It is not necessary to preclude 100% of a signal from getting through in order to cause the signal to be unrecognizable. As mentioned, typically, if one-third or so of an analog voice transmission is disrupted, enough is missing to preclude understanding the message and it sounds garbled. Also as discussed above, if 1 bit out of 10 can be disrupted in tactical digital communication systems, communication can be adequately disrupted. Therefore, methods have been devised to rapidly switch an exciter from one frequency to another fast enough so that three analog signals can be disrupted with the same jammer hardware. Against digital signals, perhaps up to 10 targets can be simultaneously precluded from communication. Even with time-sharing, though, look-through techniques are still required to maximize the utility of the jammer. If the output of the power amplifier is tuned to maximize the power transferred to the antenna, it must be rapidly switched as the exciter changes frequency. Rapidly retuning high-power filters is difficult at best.

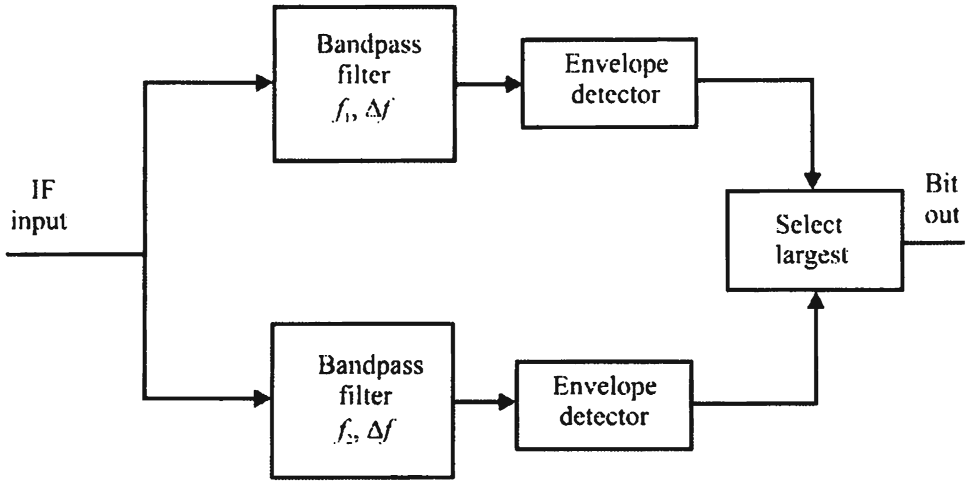

Bandpass filter f_1, Δf
Envelope detector
IF input
Select largest
Bit out
Bandpass filter f_2, Δf
Envelope detector

Figure 3.5 Block diagram for a noncoherent detector for FSK.

3.8 Barrage Jamming

A barrage jammer is a likely candidate for a UAS-borne jammer that is used against deep, second-echelon forces. This indeed puts the jammer much closer to the adversary than friendly communications, and can essentially completely deny communications within a considerable radius of the jammer. Typically for communication screening a barrage jammer is required. This is when a broad frequency range, perhaps the entire band over which friendly communications will transpire, is transmitted from the screening jammer. The jammer must be much closer to the targeted ES/SIGINT system than the friendly communications; otherwise the jammer will interfere with those friendly communications. One of the ways to implement such a jammer is to generate a relatively narrow (say, 1 MHz) signal comprised of a carrier frequency modulated with noise. This signal then is stepped from one 1 MHz portion of the spectrum to the next, usually, but not necessarily, in succession, dwelling at each step for some period of time, for example, 1 ms. One can cover, say, the lower portion of the VHF frequency band of 30 to 90 MHz in 60 ms [15].

To see how effective such a jammer could be, consider a BFSK signal with incoherent demodulation at the receiver. More will be said later about signaling techniques but for now a BFSK signal is a digital signal that uses two tones to send binary digits. This is typical for modern-day frequency-hopping VHF communication equipment. The demodulator would look something like that shown in Figure 3.5. The two tone frequencies are given by f_1 and f_2, and they occupy a bandwidth of Δf as shown in Figure 3.6. Each channel has a bandwidth in Hz given by B_{ch} and the channel center frequency is f_{c}. The jamming waveform is stepped at a rate of $R_{\text{j}} = 1/T_{\text{j}}$. The instantaneous bandwidth of the jammer is given

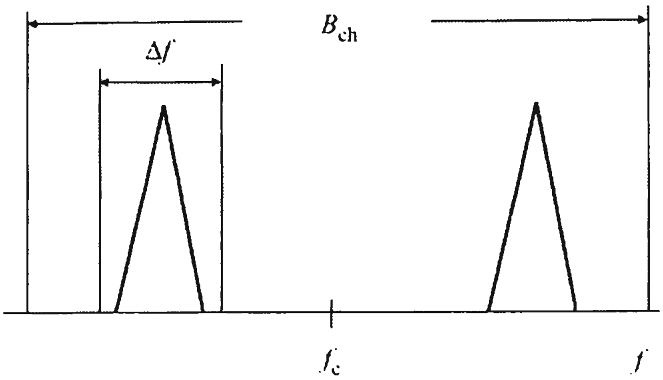


Figure 3.6 Details of the RF channel for the example. The signal is FSK, with each tone occupying a bandwidth of Δf, while the channel bandwidth is B_{ch}.

by *IBW*, which is the bandwidth of the jammer at any instant. The total bandwidth covered is given by *TBW*, which is comprised of one or more IBWs. The target frequency need not be known exactly and the epoch (timing) information of the target signal is not known. Although the outputs of the detectors in general are correlated due to imperfect filtering, it is assumed for now that these outputs are uncorrelated.

The probability of a symbol error, denoted as P_S, includes the probability of a symbol error when the jamming signal is present, .denoted as P_p, as well as the probability of a symbol error when the jammer is absent, denoted as P_a. Since these events are mutually exclusive

$$\begin{aligned} P_S =& \Pr\{\text{jammer present}\}\Pr\{\text{symbol error}|\text{jammer present}\} \\ &+\Pr\{\text{jammer absent}\}\Pr\{\text{symbol error}|\text{jammer absent}\} \end{aligned} \tag{3.2}$$

In general, the probability of a symbol error (for BFSK also a bit error) is given by [13]

$$P_S = \frac{1}{2} e^{-\frac{S}{2N_t + J_1 + J_2}} \tag{3.3}$$

where S is the power in the signal, N_t is the noise level, and the J_i's are the amounts of jammer noise that make it through the bandpass filters. In this case we will assume that $J_1 = J_2 = J$. Thus

$$\Pr_a = \Pr\{\text{symbol error}|\text{jammer absent}\} = \frac{1}{2} e^{-\frac{1}{2}\left(\frac{S}{N_t}\right)} \tag{3.4}$$

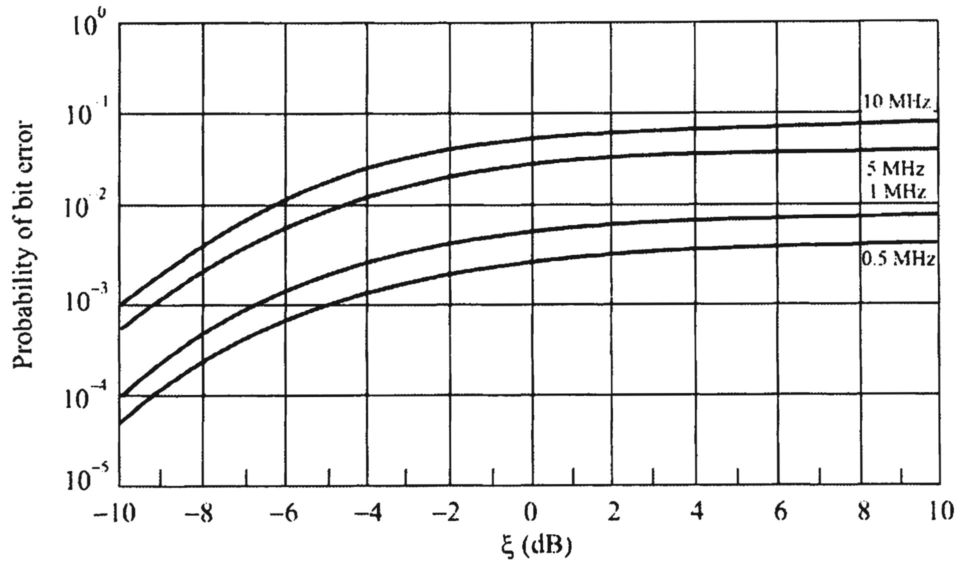


Figure 3.7 Bit error degradation caused by stepped/scanned barrage jamming as a function of the instantaneous bandwidth of the jammer and the JSR according to (3.7). (*Source:* [14]. © IEEE 1993. Reprinted with permission.)

$$\mathrm{Pr}_{\mathrm{p}} = \mathrm{Pr}\{\text{symbol error} \mid \text{jammer present}\} = \frac{1}{2} e^{-\frac{1}{2}\left(\frac{S}{N_t + J}\right)} \tag{3.5}$$

The probability of the jammer being present at any particular channel is given by

$$\mathrm{Pr}\{\text{jammer present}\} = \frac{IBW}{TBW} \tag{3.6}$$

Since Pr{jammer absent} = 1 − Pr{jammer present}, then

$$P_{\mathrm{S}} = \frac{IBW}{TBW}\mathrm{Pr}_{\mathrm{p}} + \left(1 - \frac{IBW}{TBW}\right)\mathrm{Pr}_{\mathrm{a}} \tag{3.7}$$

As an example of how well this can work, the results of calculating the probability of producing a symbol error for a variety of different instantaneous bandwidths of the jammer are shown in Figure 3.7 [14]. In this example, the SNR of the intended signal at the receiver is 20 dB. It is well known that the RF environment on a battlefield is very noisy. BERs of 10^{-2} are not uncommon (compare this to telephone quality transmissions where BERs of 10^{-6} are typical). Much of the equipment designed for tactical use is designed for a BER of 10^{-2}. Therefore, achieving BER higher than this can be viewed as successful jamming.

As seen from Figure 3.7, this can be achieved with, say, an IBW of 5 MHz, at a JSR of about –5 dB, relatively easy to achieve with a UAS-borne jammer.

Therefore, as this example shows, disruption of communications can occur for relatively low values of the J/S ratio. If a BER rate of 10^{-2} or better (1 bit or fewer out of 100 in error) is required for reliable communications, then a JSR of –6 to –4 dB, corresponding to instantaneous bandwidths of 5 MHz and 10 MHz, respectively, in this example is all that is necessary to accomplish this amount of degradation.

3.9 Follower Jammer

For LPI targets, particularly those employing frequency hopping, either a barrage jammer or a narrowband jammer can be used, but in the latter case, the ES equipment must be able to follow the transmitter as it changes frequency. This is a matter of identifying the frequency to which the transmitter moves and also identifying that the new signal belongs to the same transmitter, which is no easy task. Such a scheme is called a *follower jammer* [7, 12, 16–19].

Frequency hopping a transmitter is a jamming avoidance EP technique. The frequency of the transmitter, as well as the tuned frequency of the receiver, is changed rapidly so that a jammer, fixed on a frequency, has minimum effect on the whole transmission. Follower jamming is a way to combat this EP technique, albeit at a fairly high price in EA system complexity (the ES part of the system actually becomes substantially more complex).

The concept of a follower jammer is not new. It has long been standard *communication electronics operating instructions* (CEOI)—the standard operating procedure for tactical communications—practice that, if jammed at one frequency, change to a backup frequency that is not being jammed. This is a form of (very) slow frequency hopping. Tactical jammers have had to track such changes in frequency ever since jammers have been invented.

Since frequency-hopped radios typically use BFSK as the modulation, a tone jammer can only cover one or the other of the BFSK frequencies. For this reason, modulation is usually added to such jammers so that the whole communication channel is jammed, not just one of the tones. Frequently this modulation consists of a noise signal frequency modulated onto the carrier.

3.10 Jamming LPI Targets

In the case of jammers that operate against frequency-hopping targets, we can either use a follower jammer or a barrage jammer. A follower jammer must detect

In the case of jammers that operate against frequency-hopping targets, we can either use a follower jammer or a barrage jammer. A follower jammer must detect the frequency of the target, identify that as the target of interest, and then apply narrowband jamming power to that frequency. The problem with barrage jammers is, as mentioned, unintentional fratricide [7, 16–29].

Jamming frequency-hopped targets is dependent on the relative distances between the transmitter and the receiver, the jammer and the receiver, and the jammer and the transmitter. If the jammer is too far away from the transmitter relative to the distance between the transmitter and the receiver, the signal will arrive at the jammer after it has already been received at the receiver. The hopping communication system will have already moved to its next frequency, rendering jamming ineffective. The same problem occurs if the distance between the jammer and the receiver relative to the distance between the transmitter and the receiver is too large. Even though the jammer receives the signal in time to ascertain that it is the correct target, the signal emitted from the jammer must travel too far to reach the receiver in time to prevent communication, again rendering jamming ineffective.

The speed at which a frequency-hopped communication network hops is also a consideration for effective EA against it. Since the propagation velocity of the signals is a constant, given a specific configuration of the network and jammer, as the hop rate is increased, a point is reached where the transmitter and receiver hop to the next frequency just as the jamming signal arrives at the receiver at the old frequency, rendering EA ineffective. For that scenario and all faster hop rates the jamming will be ineffective. The closer the jammer is to the receiver and/or transmitter, the faster this rate must be to render the jamming ineffective, but for all scenarios there is such a rate. It is for these reasons that it is important to deploy EA assets as close to the communicating nodes as possible.

3.11 Concluding Remarks

Successfully denying an adversary the use of the communication spectrum is the purpose of communication EW. It depends on many factors, including the link distances between the transmitter and the receiver relative to the link distance between the jammer and the receiver, the ERP from the jammer relative to the ERP of the transmitter, and the type of communications involved.

Digital communications are easier to jam than analog forms. Creating the necessary BER to deny such communications requires less energy and can be done much more safely than before.

Jamming LPI communications is not difficult, especially if digital modulations are involved. What is difficult with these forms of communication is battlespace is difficult. The alternative to surgically jamming a signal such as this is to barrage jam a significant portion of the RF spectrum. This can only be done, however, in situations where friendly communications are not affected.

References

[1] Waltz, E., *Information Warfare Principles and Operations*, Norwood, MA: Artech House, 1998, pp. 165–168.

[2] Schleher, D. C., *Electronic Warfare in the Information Age*, Norwood, MA: Artech House, 1999.

[3] Adamy, D., *EW 101: A First Course in Electronic Warfare*, Norwood, MA: Artech House, 2001.

[4] Neri, F., *Introduction to Electronic Defense Systems*, Norwood, MA: Artech House, 1991.

[5] Mosinski, J. D., "Electronic Countermeasures," *Proceedings of the IEEE Tactical Communications Conference*, *Vol. 1*, 28–30 April 1992, pp. 191–195.

[6] Torrieri, D. J., *Principles of Secure Communication Systems*, 2nd Ed., Norwood, MA: Artech House, 1992, pp. 275–288.

[7] Torrieri, D. J., "Fundamental Limitations on Repeater Jamming of Frequency-Hopping Communications," *IEEE Journal on Selected Areas of Communications*, Vol. SAC-7, No. 5, May 1989.

[8] Lathi, B. P., *Communication Systems*, New York: John Wiley & Sons, 1968.

[9] Gagliardi, R. M., *Introduction to Communications Engineering*, 2nd Ed., New York: Wiley, 1988.

[10] Mosinski, J. D., "Electronic Countermeasures," *Proceedings IEEE MILCOM Conference*, 1992, p. 193.

[11] Proakis, J. G., *Digital Communications*, New York: McGraw-Hill, 1995.

[12] Simon, M. K., et al., *Spread Spectrum Communications Handbook*, New York: McGraw-Hill, 1994.

[13] Torrerri, D. J., *Principles of Secure Communication Systems*, 2nd Ed., Norwood, MA: Artech House, 1992, p. 19.

[14] Poisel, R. A., "Performance Analysis of a Stepped/Scanned Barrage Jammer," *Proceedings MILCOM 1993*, Boston, MA, 1993.

[15] Peterson, R. L., R. E. Ziemer, and D. E. Borth, *Introduction to Spread Spectrum Communications*, Upper Saddle River, NJ: Prentice Hall, 1995.

[16] Simon, M. K., "The Performance of M-ary FH-DPSK in the Presence of Partial-Band Multitone Jamming," *IEEE Transactions on Communication*, Vol. COM-30, No. 5, May 1982, pp. 953–958.

[17] Torrieri, D. J., *Principles of Secure Communication Systems*, Norwood, MA: Artech House, 1992, pp. 245–257.

[18] Putnam, C. A., S. S. Rappaport, and D. L. Schilling, "Tracking of Frequency-Hopped Spread Spectrum Signals in Adverse Environments," *IEEE Transactions Communications*, Vol. COM-31, No. 9, August 1983.

[19] Hassan, A. A., W. E. Stark, and J. E. Hershey, "Frequency-Hopped Spread Spectrum in the Presence of a Follower Partial-Band Jammer," *IEEE Transactions on Communications*, Vol. 41, No. 7, July 1993, pp. 1125–1131.

[20] Simon, M. K., and A. Polydoros, "Coherent Detection of Frequency-Hopped Quadrature Modulations in the Presence of Jamming – Part I: QPSK and QASK Modulations," *IEEE Transactions on Communications*, Vol. COM-29, No. 11, November 1981, pp. 1644–1660.

[20] Simon, M. K., and A. Polydoros, "Coherent Detection of Frequency-Hopped Quadrature Modulations in the Presence of Jamming – Part I: QPSK and QASK Modulations," *IEEE Transactions on Communications*, Vol. COM-29, No. 11, November 1981, pp. 1644–1660.

[21] Simon, M. K., "Coherent Detection of Frequency-Hopped Quadrature Modulations in the Presence of Jamming – Part II: QPR Class I Modulations," *IEEE Transactions on Communications*, Vol. COM-29, No. 11, November 1981, pp. 1661–1668.

[22] Simon, M. K., G. K. Huth, and A. Polydoros, "Differentially Coherent Detection of QASK for Frequency Hopping Systems – Part I: Performance in the Presence of Gaussian Noise Environment," *IEEE Transactions on Communications*, Vol. COM-30, No. 1, January 1982, pp. 158–164.

[23] Simon, M. K., "Differentially Coherent Detection of QASK for Frequency-Hopping Systems – Part II: Performance in the Presence of Jamming," *IEEE Transactions on Communications*, Vol. COM-30, No. 1, January 1982, pp. 165–172.

[24] Lee, J. S., L. E. Miller, and Y. K. Kim, "Probability of Error Analysis of a BFSK Frequency-Hopping System with Diversity under Partial-Band Jamming Interference–Part II: Performance of Square-Law Nonlinear Combining Soft Decision Receivers," *IEEE Transactions on Communications*, Vol. COM-32, No. 12, December 1984.

[25] Milstein, L. B., and D. L. Schilling, "The Effect of Frequency-Selective Fading on a Noncoherent FH-FSK System Operating with Partial-Band Tone Interference," *IEEE Transactions on Communications*, Vol. COM-30, No. 5, May 1982, pp. 904–912.

[26] Miller, L. E., et al., "Analysis of an Antijam FH Acquisition Scheme," *IEEE Transactions on Communications*, Vol. 40, No. 1, January 1992, pp. 160–170.

[27] Kwon, H. M., L. E. Miller, and J. S. Lee, "Evaluation of a Partial-Band Jammer with Gaussian-Shaped Spectrum Against FH/MFSK," *IEEE Transactions on Communications*, Vol. 38, No. 7, July 1990, pp. 1045–1049.

[28] Viswanathan, R., and S. C. Gupta, "Performance Comparison of Likelihood, Hard-Limited, and Linear Combining Receivers for FH-MFSK Mobile Radio-Base-to-Mobile Transmissions," *IEEE Transactions on Communications*, Vol. COM-11, No. 5, May 1983, pp. 670–903.

[29] Hassan, A. A., J. E. Hershey, and J. E. Schroeder, "On a Follower Tone-Jammer Countermeasure Technique," *IEEE Transactions on Communications*, Vol. 43, No. 2/3/4, February/March/April 1995, pp. 754–756.

Part I – Basics

Chapter 4

Electromagnetic Signal Propagation

4.1 Introduction

Before describing the characteristics of EW systems it is important to understand some of the fundamental properties of propagating signals, or EM waves. For those familiar with this topic, this chapter can be skipped. For those readers interested in further reading, [1–7] are recommended.

The modes of signal propagation can be dependent on the frequency of the signal, while some of the modes are independent of the frequency. These characteristics are pointed out in the appropriate following discussions.

The design of the RF front ends, especially the antenna, of EW systems depends to a great extent upon how signals propagate. The fundamental equations that govern the movement of signals from a transmitter, through space, and at a receiver are explained in this chapter. In this case, "signals" refers to RF signals, which for our purposes are assumed to start around 500 kHz.

This chapter is organized as follows. The next section describes the fundamental tenants of electromagnetic waves, describing the electric and magnetic fields. That is followed by a presentation of the common and new designations for the frequency bands. Then EM wave polarization is defined. The wave quantity entitled power density is then derived. Free-space propagation characteristics are presented next. That is followed by several sections that discuss the propagation modes prevalent in the higher frequency bands including: ground wave with the direct wave free-space power loss model and surface wave; wave diffraction around corners; wave reflections including derivation of the 2-ray reflection model; ducting; meteor burst; and scattering. Then a section on the introduction to signal fading phenomena in the higher frequency bands is presented. This section is just an introduction since there is considerably more discussion later about fading. That is followed by a summarization of the mobile VHF channel. The last section presents a discussion of signal propagation in the lower frequency ranges where the ionosphere plays a significant role.

4.2 Signal Propagation

A signal generated in a transmitter leaves that transmitter at a specified power level and is sent to an associated antenna usually via interconnecting cables. The antenna typically has a gain, which increases the level of the signal in certain preferred directions. As the signal propagates through the atmosphere, it suffers losses due to the spreading of the signal in space and losses due to encountered obstacles. It arrives at a receiver antenna at some power level, which usually has some characteristic gain, thus increasing the level of the signal. This signal is then presented to the receiver from the receiver antenna.

An RF signal propagates through space at about the speed of light, 3×10^8 m/s, so, with f expressed in MHz, its wavelength is given by

$$\lambda = \frac{300}{f} \tag{4.1}$$

At a distance greater than about 1 wavelength from the transmitter the RF signal is composed of E and H fields which are orthogonal and propagate in a direction at right angles to the E and H fields as illustrated in Figure 4.1.

Neither the electric field nor the magnetic field can exist without the other. The electric field has units of volts per meter while the magnetic field has units of amperes per meter.

A signal radiates from an *isotropic* antenna (approximated by a point source) in an ever-expanding sphere, until it encounters something that perturbs that sphere. (An isotropic antenna is a theoretical antenna that radiates equally in all directions. While hypothetical, it is frequently used for comparison purposes.) At a significant enough distance from the transmitter, typically taken to be at least 10 times the wavelength, the spherical wave front is frequently approximated as a plane over the dimensions of most antennas.

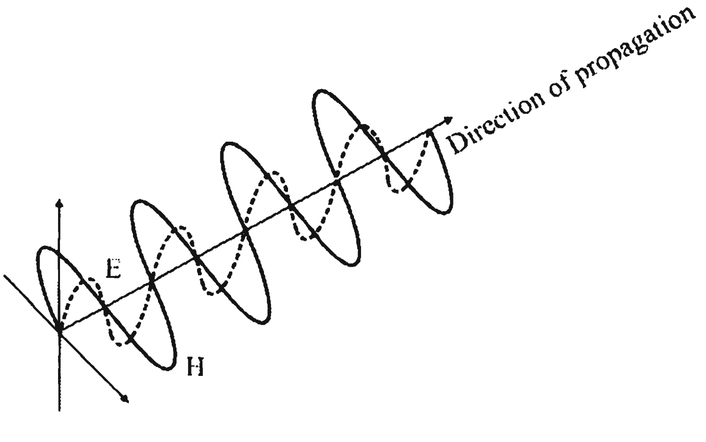


Figure 4.1 E and H fields associated with an electromagnetic wave.

There are several modes of signal propagation. The major modes that are of importance for communication EW system design are: ground wave consisting of the direct wave and surface wave, reflected wave, refracted wave, diffracted wave, and scatter wave. For ground-to-ground communications, useful VHF and above signal propagation is limited to the tropospheric layer of the atmosphere, which ranges in altitude from 9 km at the Earth's poles to about 17 km at the equator [8]. On the other hand, HF signal propagation phenomena takes advantage of the ionosphere for long-distance communication. For ground-to-air or air-to-air communications, the direct wave is the method most frequently used. Of course, ground-to-satellite communications are direct.

4.3 Radio Frequency Band Designations

The radio frequency spectrum is divided into designated bands with the common designations shown in Table 4.1. Communication services of some sort are provided in virtually all of the RF bands. The frequency bands are shown graphically in Figure 4.2.

The microwave band (300 MHz to 300 GHz) has long been subdivided into sub-bands. It has recently undergone a change in sub-band designation, however. The bands, along with their old and new designations, are given in Table 4.2.

Table 4.1 Frequency Band Designations

Frequency Band	Name	Designation
3–30 kHz	Very Low Frequency	VLF
30–300 kHz	Low Frequency	LF
300–3,000 kHz	Medium Frequency	MF
3–30 MHz	High Frequency	HF
30–300 MHz	Very High Frequency	VHF
300–3,000 MHz	Ultra High Frequency	UHF
3–30 GHz	Super High Frequency	SHF
30–300 GHz	Extra High Frequency	EHF
300–3,000 GHz	Optical	

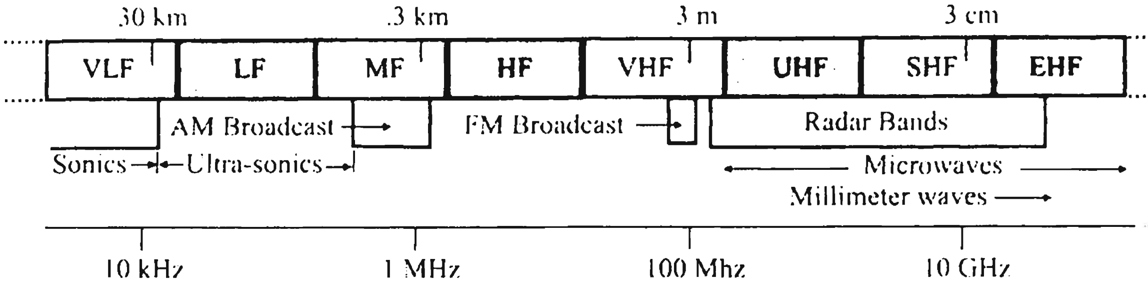


Figure 4.2 Summary of frequency bands. The wavelengths are listed on the top while the frequencies are shown on the bottom.

4.4 Polarization

The *polarization* of an EM wave is the orientation of the electric field component of the wave. Although the polarization of an EM wave can be anything, the most common forms of polarization are linear (vertical or horizontal, where these orientations are normally relative to the surface of the Earth), circular, and elliptical. In these latter two categories the electric field is intentionally rotated as the signal traverses space. Manipulating the phase of the signal at the transmitter creates this rotation.

Neglecting the impacts of the environment through which the wave propagates, the polarization of an EM wave is determined by the transmitting antenna. Antennas can be designed to impose any of the polarizations on waves. Neglecting the effects of the environment is an unaffordable luxury, however.

Table 4.2 New Designations of the Higher Frequency Bands

Frequency Band	Old Designation	New Designation
.5–1 GHz	UHF	C
1–2 GHz	L	D
2–3 GHz	S	E
3–4 GHz	S	F
4–6 GHz	C	G
6–8 GHz	C	H
8–10 GHz	X	J
10–14.4 GHz	X	J
14.4–18 GHz	Ku	J
18–20 GHz	K	J
20–26.6 GHz	K	K
26.6–40 GHz	Ka	K

Metallic objects off of which a wave may reflect change the polarization. The ionosphere changes the polarization of HF signals refracted by it. Reflections of a wave off the Earth's surface can change the polarization. These are but a few of the causes of polarization distortion.

4.5 Power Density

The *power density* of an EM wave is a measure of the power per unit area in the wave at any point in space and is a vector. Herein it is denoted as $\vec{P}_{\text{den}}$ and is given in units of watts/meter^2. The *field strength* of an EM wave, *E*, is a measure of the voltage potential differences in the wave as a function of distance and is given in units of volts/meter. Although it is not necessary, herein only sinusoidal waves will be considered unless otherwise indicated. That allows the phasor forms of the signals to be used. As shown in Figure 4.1, the *E* and *H* waves can be considered as vectors, which have both an amplitude and a direction. As vectors they are denoted by $\vec{E}$ and $\vec{H}$. According to the *Poynting Theorem* power density vector is given by

$$\vec{P}_{\text{den}} = \frac{1}{2}\text{Re}\left(\vec{E} \times \vec{H}^*\right) \tag{4.2}$$

where × represents the vector cross product and $\vec{E}$ and $\vec{H}$ are the peak field values. The power density can also be represented as

$$\vec{P}_{\text{den}} = P_{\text{den}}\vec{u}_{\text{r}} \tag{4.3}$$

where P_{den} is the magnitude of the vector and $\vec{u}_{\text{r}}$ represents the unit vector in the direction of propagation.

When dealing with $\vec{H}$ or $\vec{E}$, care must be taken to know whether the peak or RMS value (often called average) is being used. Most textbooks on antennas use peak values and most regulations and technical articles on antennas use RMS values. In order to be consistent herein, RMS values will be used unless otherwise noted.

Therefore, if the electric or magnetic field strength at any point is known, the power density can be calculated. This will be convenient when we want to know how much power the transmitter can put into a particular point in space. Let

$$\vec{E} = E_{\text{peak}} \cos(2\pi f t)\vec{u}_{\text{E}} \tag{4.4}$$

$$\vec{H} = H_{\text{peak}} \cos(2\pi f t)\vec{u}_{\text{H}} \tag{4.5}$$

where $\vec{u}_{\text{E}}$ and $\vec{u}_{\text{H}}$ are unit vectors in the direction of $\vec{E}$ and $\vec{H}$, respectively. The *characteristic impedance*, Z_0, of the propagation medium is given by the ratio of the magnitudes of the electric wave to the magnetic wave. Thus

$$Z_0 = \frac{E_{\text{peak}}}{H_{\text{peak}}} = \sqrt{\frac{\mu}{\varepsilon}} = 120\pi \approx 377\Omega \tag{4.6}$$

where $\mu = 4\pi \times 10^{7}$ is the permeability of free-space and $\varepsilon = 8.854 \times 10^{-12}$ is the permittivity of free-space. Therefore, since $\vec{u}_{\text{E}}$ and $\vec{u}_{\text{H}}$ are orthogonal,

$$\begin{aligned} P_{\text{den}} &= \frac{1}{2}\text{Re}\left(E_{\text{peak}} \frac{E_{\text{peak}}}{Z_0} \right) \\ &= \frac{1}{2}\frac{E_{\text{peak}}^2}{Z_0} \end{aligned} \tag{4.7}$$

However, since for sinusoidal waves, $E_{\text{rms}} = E_{\text{peak}}/\sqrt{2}$,

$$\begin{aligned} P_{\text{den}} &= \frac{E_{\text{rms}}^2}{Z_0} \\ &= \frac{E_{\text{rms}}^2}{120\pi} \end{aligned} \tag{4.8}$$

Therefore, the electric field strength is related to the power density by

$$E_{\text{rms}} = \sqrt{120\pi P_{\text{den}}} \tag{4.9}$$

From these equations the electric field strength can be calculated versus distance between the transmitter and receiver.

The *sensitivity* of receiving systems is often given in terms of microvolts per meter (μV/m), that is, in terms of field strength. This is a system level specification in that it reflects the degree to which the system can extract EM wave energy. It best describes the ability of the system antenna to deliver signal power to the remainder of the system. Typical values are 1–5 μV/m. Equivalently, sensitivity is sometimes specified in terms of decibels relative to 1 μV/m. Since a decibel is a measure of power, conversion to power terms is necessary to change such a specification to absolute values. Recall that decibel is defined as

$$\mathrm{dB} = 10\log\frac{P_2}{P_1} \tag{4.10}$$

When specifying decibels relative to 1 μV/m, the two power values are calculated assuming the same impedance value, usually the characteristic impedance of free-space, Z_0. Thus

$$\begin{aligned}\mathrm{dB}_{1\mu\mathrm{V/m}} &= 10\log\frac{(\mathrm{V/m})^2 / Z_0}{(1\mu\mathrm{V/m})^2 / Z_0}\\ &= 20\log\frac{\mathrm{V/m}}{1\mu\mathrm{V/m}}\end{aligned} \tag{4.11}$$

so

$$\mathrm{V/m} = 10^{-6} \times 10^{\mathrm{dB}_{\mu\mathrm{V/m}}/20}$$

(of course the 10^{-6} is not used if the results are desired in units of μV/m).

4.6 Free-Space Propagation

Free-space propagation refers to the propagation mode between two antennas where there is no obstacle between the antennas to interfere with the ever-expanding spherical surface of the signal emitted from the transmitting antenna, when the transmitting antenna is isotropic. Free-space propagation is only possible in outer space where there is relatively little matter. Some air-to-air, including satellite, communications approximate free-space situations fairly well, however.

At a distance *r* from an isotropic transmit antenna the amount of EM wave power that passes through a unit area on the surface of the spherical surface is, by simple calculus,

$$P_{\mathrm{den}} = \frac{P_{\mathrm{transmitted}}}{4\pi r^2} \tag{4.12}$$

where $P_{\text{transmitted}}$ is the amount of power emitted by the transmitter. The EM must be in the far field in order for (4.12) to apply.

The amount of power received by the receiver is given by the amount of this power density absorbed by the receiver antenna. The effective area, A_{eff}, determines this. Therefore,

$$P_R = P_{\text{den}} A_{\text{eff}} \tag{4.13}$$

The effective area of an isotropic receiver antenna is given by

$$A_{\text{eff}} = \frac{\lambda^2}{4\pi} \tag{4.14}$$

Therefore the power received in free-space propagation between two isotropic antennas is given by

$$\begin{aligned} P_{\text{received}} &= \frac{P_{\text{transmitted}}}{4\pi r^2} \frac{\lambda^2}{4\pi} \\ &= P_{\text{transmitted}} \left(\frac{\lambda}{4\pi r} \right)^2 \end{aligned} \tag{4.15}$$

The *gain* of an antenna is a measure of how particular directions are favored for propagation over others. More will be said about gain later. Suffice it here to say that the amount of power transmitted is given by

$$P_{\text{transmitted}} = G_T P_T \tag{4.16}$$

where P_T is the amount of signal power accepted by the antenna and G_T is the transmit antenna gain in some direction. A similar relationship exists for the receiver antenna,

$$P_R = G_R P_{\text{received}} \tag{4.17}$$

where G_R is the receiver antenna gain in some direction and P_R is the power available from the antenna.

From above,

$$P_{\text{den}} = \frac{G_{\text{TR}} P_{\text{T}}}{4\pi r^2} = \frac{E_{\text{rms}}^2}{120\pi} \tag{4.18}$$

so the free-space RMS field strength can be calculated to be

$$E_{\text{rms}} = \frac{\sqrt{30 G_{\text{TR}} P_{\text{T}}}}{r} \tag{4.19}$$

Define the *free-space path loss* as the ratio of the power out of the receiver antenna to the power input to the transmit antenna

$$L = \frac{P_{\text{R}}}{P_{\text{T}}} = G_{\text{TR}} G_{\text{RT}} \frac{\lambda^2}{(4\pi r)^2} \tag{4.20}$$

This is known as *Friis' expression* for free-space path loss. In dB it is

$$L_{\text{dB}} = \begin{Bmatrix} -32.2 \\ -36.6 \end{Bmatrix} - 20\log(f_{\text{MHz}}) - 20\log(r) + G_{\text{TR,dB}} + G_{\text{RT,dB}}, \begin{Bmatrix} \text{km} \\ \text{mile} \end{Bmatrix} \tag{4.21}$$

when G_{TR} and G_{RT} are expressed in dB. Thus, again in dB, when P_{T} is expressed in decibels relative to 1W

$$P_{\text{R}} = P_{\text{T}} - L_{\text{dB}} \tag{4.22}$$

or, when P_{T} is expressed in decibels relative to 1 mW

$$P_{\text{R}} = P_{\text{T}} - L_{\text{dB}} + 30 \tag{4.23}$$

This expression ignores the cable losses associated with the transmitter and receiver. Denoting these by L_{T} and L_{R}, when they are included the expressions become

$$P_{\text{R}} = P_{\text{T}} - L_{\text{dB}} - L_{\text{T}} - L_{\text{R}} \tag{4.24}$$

and

Table 4.3 Specific Loss of Some Common Cable Types

	Loss Per Foot (dB)		
Cable Type	**At 100 MHz**	**At 400 MHz**	**At 1000 MHz**
RG6/U	0.019	0.043	0.065
RG58/U	0.05	0.11	0.2
RG59/U	0.038	0.075	0.11
RG8/U	0.025	0.054	0.092
RG174/U	0.11	0.22	0.32
RG188/U	0.105	0.18	0.3
RG213/U	0.025	0.055	0.095

$$P_\text{R} = P_\text{T} - L_\text{dB} - L_\text{T} - L_\text{R} + 30 \tag{4.25}$$

respectively. Specific values for L_T and L_R for some common types of cable are given in Table 4.3. These loss values correspond to newly manufactured cable, and the actual loss will depend on installation parameters, age of the cable, temperature, etc.

4.7 Ground Wave

The *ground wave* is a wave that propagates from the transmitter of the receiver close to or next to the ground. It consists of two different components: the direct wave and the surface wave.

4.7.1 Direct Wave

If there is line of sight between the transmitter and the receiver, then the *direct wave* carries energy between the transmit antenna and receiver antenna in a direct route. As above, at sufficiently large distances above the Earth's surface, the free-space propagation model may apply to the direct wave. Closer to the Earth, however, that is not the case. In fact, as illustrated in Figure 4.3, the amount of power available from the receiver antenna is given by

$$P_\text{R} = P_\text{T} \frac{G_\text{TR} G_\text{RT} \lambda^2}{(4\pi)^2 r^n} \tag{4.26}$$

where P_T = power input to the transmit antenna from the cable connecting the transmitter to the transmit antenna (watts); G_TR = gain of the transmit antenna in the direction of the receiver antenna (unit-less); G_RT = gain of the receiver antenna

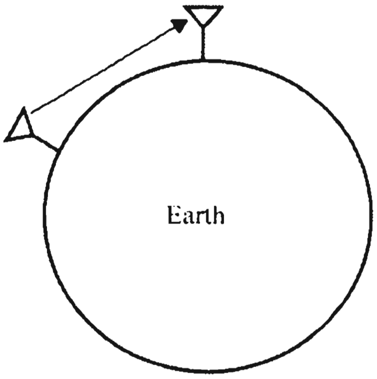


Figure 4.3 A direct wave travels in a straight line from the transmit antenna to the receiver antenna.

in the direction of the transmit antenna (unit-less); λ = wavelength of the signal (meters); r = distance between the transmitter and the receiver (meters); and n is discussed below.

The electric field strength versus distance for radios with an ERP of 40 and 50 dBm is shown in Figure 4.4. This chart assumes that the propagation exponent $n = 2$, corresponding to the free-space conditions discussed in Section 4.6. Figure 4.5 shows the electric field strength versus distance when the propagation exponent $n = 4$. Note the change in range on the abscissa. Clearly the signals fall off much more rapidly close to the Earth than high in the air. Typical specifications for electronic support systems are a sensitivity of 1 to 5 μV/m. Therefore, the signals from these radios at low altitudes ($n = 4$) can be detected at ranges less than 10 km for ground ES systems. An airborne system, on the other hand, has an intercept range exceeding 150 km for air-to-air intercept with such targets. Thus airborne systems typically have a substantial advantage for signal intercept.

Unavoidable electronic losses occur at the transmitter and receiver. These losses include cable-resistive loss, frequently expressed as I^2R losses because the loss increases as the current increases. Another source of signal loss at receiving antennas is due to mismatches in the polarization between the signal and the orientation of the antenna. Theoretically no signal is received at all if the signal is vertically polarized and the antenna is horizontally polarized or vice versa. In practice, some signal is received even in this case, however. Herein, these losses are included in the values of G_{TR} and G_{RT}.

There is a propagation model methodology based on the free-space model that is sometimes called the R^n model [9]. The mean path loss at a distance r from the transmitter is calculated as

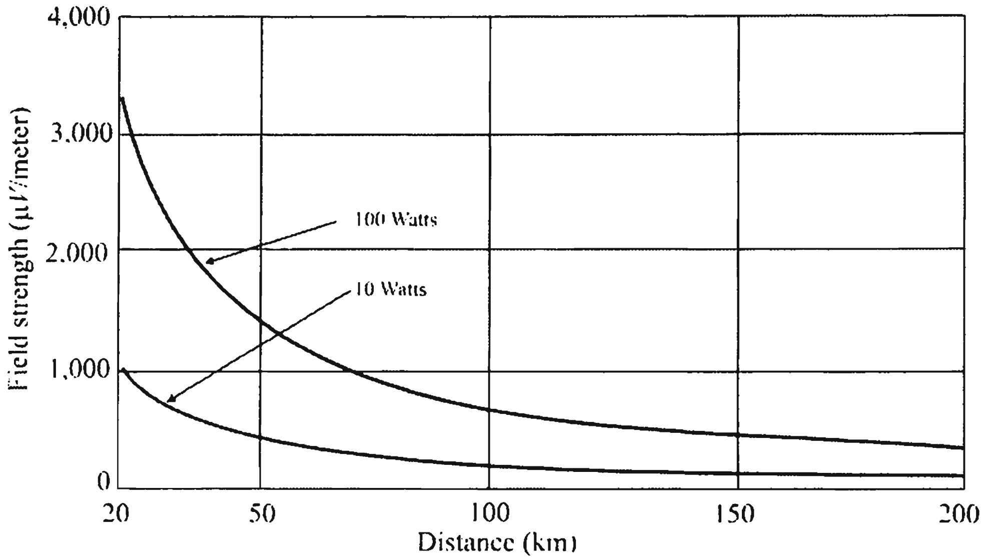


Figure 4.4 Electric field strength as a function of range for two emitter powers, where the propagation exponent $n = 3$.

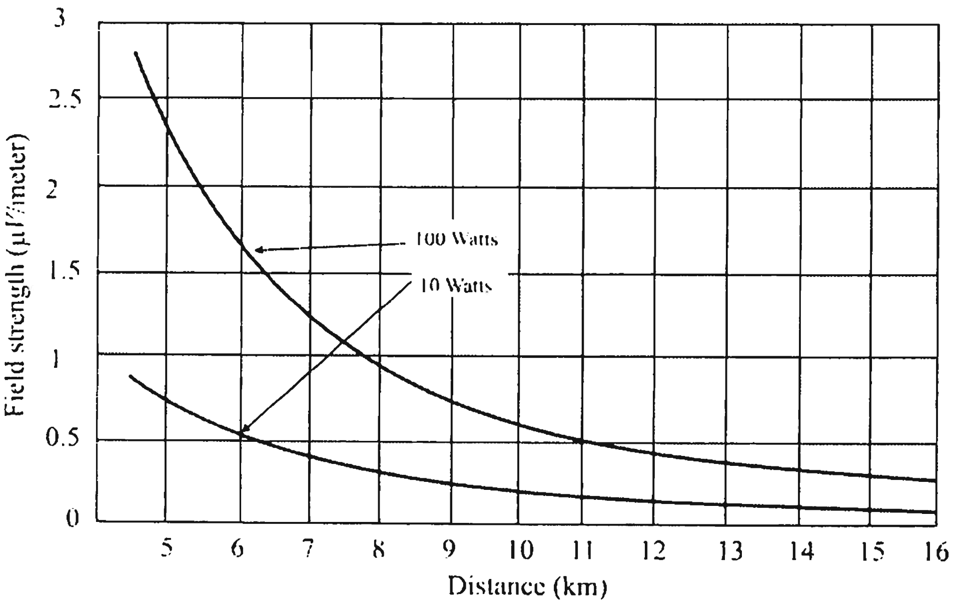


Figure 4.5 Electric field strength as a function of range for two emitter powers, where the propagation exponent $n = 4$.

$$L(r) = L(r_0) + 10n\log_{10}\left(\frac{r}{r_0}\right) \tag{4.27}$$

where r_0 is a suitably defined reference distance, frequently taken as 1 km for outdoor propagation conditions and 1m for indoor applications. $L(r_0)$ represents the path loss at the reference distance. It is frequently measured, but if it is not otherwise known, it can be estimated by

$$L(r_0) \approx 20\log_{10}\left(\frac{4\pi r_0}{\lambda}\right) \tag{4.28}$$

The parameter n is discussed below and depends on the environmental conditions. The received power, then at range r is, as above,

$$P_\mathrm{R}(r) = \frac{P_\mathrm{T} G_\mathrm{TR} G_\mathrm{RT}}{L(r)} \tag{4.29}$$

Examination of (4.29) reveals several items of note about direct-wave signal propagation. First, lower frequencies propagate better than higher ones ($\lambda = c/f$). Second, the level of the received signal decreases as a power n of the distance between the transmitter and the receiver. Away from the surface of the Earth, $n = 4$. Closer to the Earth's surface, the signal strength decreases faster than $1/r^2$, and $n = 4$ is often used. The actual exponent on r varies with each situation. It ranges from 2 to 16 or more. Therefore it can be concluded that if an antenna is raised higher into the air, better signal propagation should result. In general this is true, but to get dramatically improved performance the antennas need to be substantially elevated.

The propagation loss at any range r is a statistical parameter. As such, the loss given above is an average. In many cases, the statistics follow a lognormal distribution. Thus, there is an associated standard deviation of measurements as well. The loss exponent n is given in Table 4.4 for various propagation conditions. For outdoor environments, the standard deviation is in the range of 8 to 14 dB. On the other hand, the propagation loss exponent n and standard deviation σ are given in Table 4.5 for PCSs in several indoor conditions [9].

Most lower-echelon tactical battlefield radios used for real-time command and control of tactical forces have power levels into their associated antennas of 10–100W (40–50 dBm). The typical antenna for these configurations is a tuned whip, which, if it's in good condition, has about 2 dB of gain. Most realistic

Table 4.4 Loss Exponent n for Various Conditions

Condition	Loss Exponent, n
Free-space	2
Urban area cellular, PCS	4.0 to 4.7
Shadowed urban cellular, PCS	3 to 5
In building line of sight	1.6 to 1.8
Obstructed in building	4 to 6
Obstructed in factories	2 to 3

Source: [9].

Table 4.5 Propagation Loss Exponent n and Associated Standard Deviation σ for Several Indoor Conditions

Conditions	Frequency (MHz)	n	σ (dB)
Indoor – Retail Store	914	4.2	8.7
Indoor – Grocery Store	914	1.8	5.2
Indoor – Hard Partition Office	900	3.0	7.0
Indoor – Soft Partition Office	900	4.4	9.6
Indoor – Soft Partition Office	1,900	4.6	14.1
Indoor – Factory (LOS)	1,300	1.6–4.0	3.0–5.8
Indoor – Factory (LOS)	4,000	4.1	7.0
Indoor – Suburban Home	900	3.0	7.0
Indoor – Factory (Obstructed)	1,300	3.3	6.8
Indoor – Factory (Obstructed)	4,000	4.1	9.7
Indoor – Office Same Floor	914	3.27–4.76	5.2–14.9
Indoor – Office Entire Building	914	3.54–4.33	13.3–14.8
Indoor – Office Wing	914	4.01–4.68	4.4–8.1
Indoor – Average	914	3.14	16.3
Indoor – Through One Floor	914	4.19	5.1
Indoor – Through Two Floors	914	5.04	6.5
Indoor – Through Three Floors	914	5.22	6.7

Source: [9].

scenarios, however, do not involve perfect hardware implementations. We will assume here that the net antenna gain, cable loss, and so forth. of these radios is a net 0 dB (no gain or loss).

The so-called *range to horizon* is the distance beyond which one no longer has visual *line of sight* (LOS) between the transmitter and receiver. Assuming that the Earth is a smooth sphere, the range to horizon can be calculated according to the equation obtained from simple geometric principles as

$$r_{\mathrm{h}}^2 = (R_{\mathrm{e}} + h)^2 - r^2 \tag{4.30}$$

or

$$r_{\mathrm{h}} = \sqrt{2R_{\mathrm{e}}h + h^2} \tag{4.31}$$

where R_{e} is the radius of the Earth in km, h is the elevation of the transmitter in m, and r_{h} is the range to horizon in km. Frequently h^2 can be neglected relative to R_{e} so

$$r_{\mathrm{h}} = \sqrt{2R_{\mathrm{e}}h} \tag{4.32}$$

These distances are illustrated in Figure 4.6. Of course, here the Earth is assumed to be a sphere with an effective radius R_{e}.

Assuming that the Earth is a sphere of radius 3,960 miles, the distance to the horizon for an aircraft at 20,000 feet is 173 miles. At VHF (30–300 MHz) and above frequencies the atmosphere close to the surface of the Earth refracts (bends) radio waves that pass through it as illustrated in Figure 4.7. The net effect is to increase the effective radius of the Earth by approximately 1/3. Therefore a more correct value to use for R_{e} in (4.32) is 4/3 R_{e}, or 5,280 miles. The resultant calculation is called the *radio line of sight* (RLOS). The 4/3 factor does not always apply, however. In cases of *subrefraction*, the bending of the radio waves in the atmosphere is less than otherwise. In *super-refraction*, the waves are bent more than normal and the path length is longer than otherwise. Using similar arguments, assuming a 4/3-Earth model and a smooth Earth surface, a transmitter and receiver are within radio line of site of each other as long as they are within a distance

$$d = \sqrt{2h_{\mathrm{T}}} + \sqrt{2h_{\mathrm{R}}} \tag{4.33}$$

of each other, when d is in miles and the h_{T} and h_{R} are in feet. Thus if the transmit antenna is at a height of 20 feet and the receiver antenna is at 50 feet, the RLOS

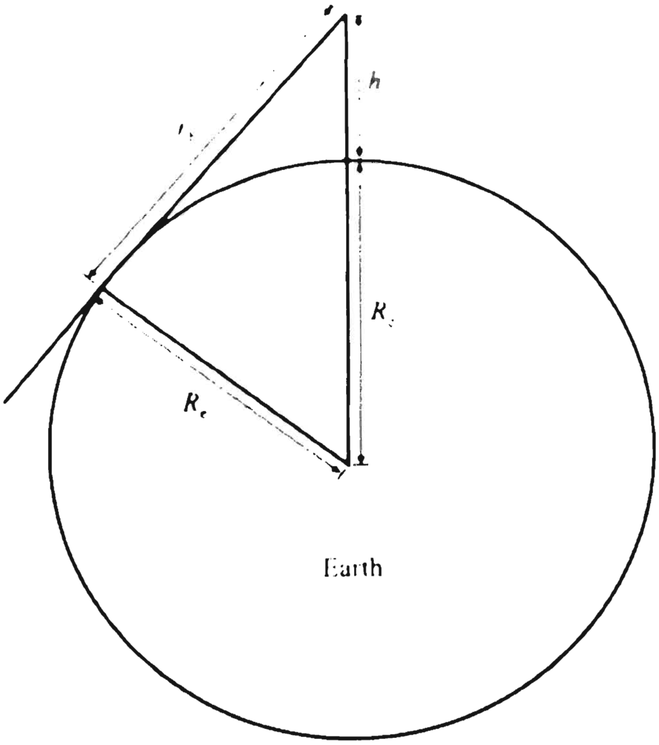


Figure 4.6 Range to horizon.

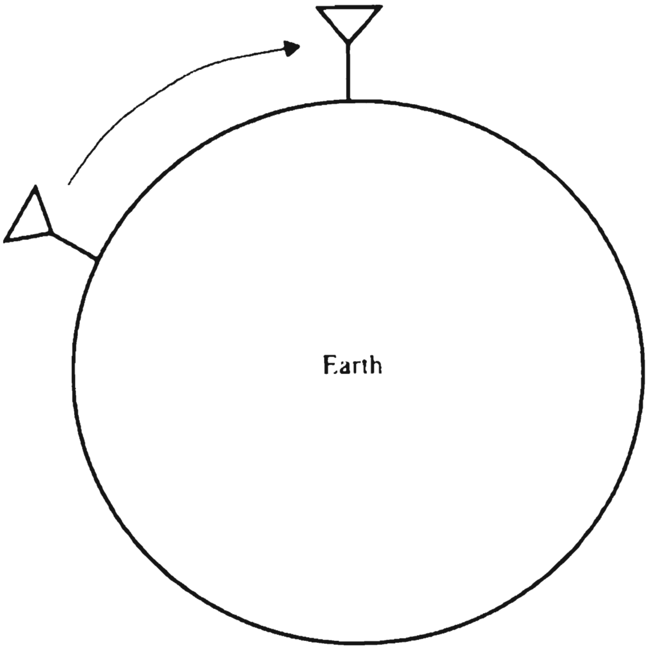


Figure 4.7 The troposphere close to the surface of the Earth refracts radio waves as they traverse through it, causing an extension of the propagation range.

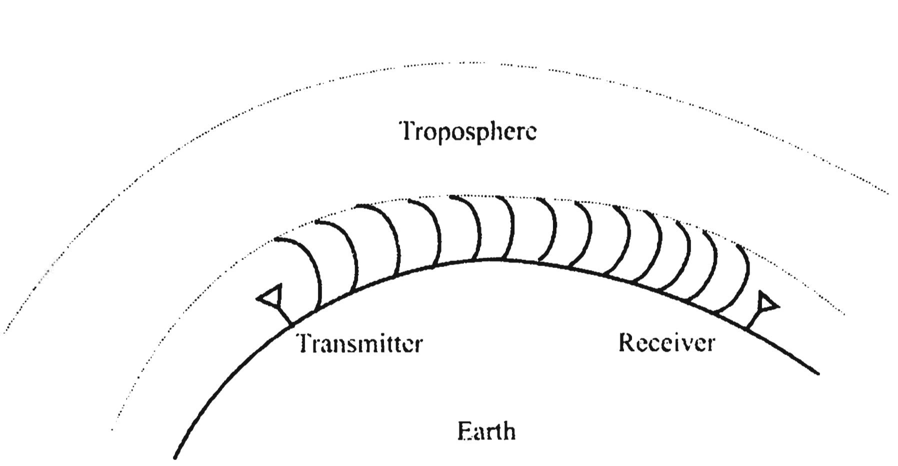


Figure 4.8 Propagation along a trough, one side of which is the Earth, is called surface-wave propagation.

is 16.3 miles. Caution should be used when applying these equations to transmitters and receivers close to the Earth's surface, however, because the Earth's surface is not smooth and, depending on the frequency, obstacles can (and frequently do) reduce this range or even preclude signal propagation at all. On the other hand, complex phenomena such as edge diffraction and reflections from surfaces can enhance signal propagation close to the Earth.

4.7.2 Surface Wave

At frequencies below approximately 50 MHz, there is a mode of propagation referred to as the *surface wave*. This wave propagates along the surface of the Earth out to considerable distances, depending on conditions, and is due to the difference in refractive index between the air and ground. At the higher frequencies the attenuation is too high to support propagation. This mode of propagation is illustrated in Figure 4.8.

As a surface wave passes over the ground, the electric field component of the wave induces a voltage in the Earth. The induced voltage causes dissipation of energy from the wave, thereby attenuating the wave as it moves along. To minimize this attenuation, the level of induced voltage must be reduced. This is accomplished by using vertically polarized waves that minimize the contact of the electric field component with the Earth. With horizontal polarization, the electric field component is parallel with the surface of the Earth and, therefore, the contact between it and the Earth is maximized. The wave is then completely attenuated within a short distance from the transmitting site. Therefore, vertical polarization is vastly superior to horizontal polarization for surface wave propagation.

The wave travels slowest at the ground which causes the wave front to tilt as illustrated in Figure 4.9. In addition, the higher the frequency the faster the wave

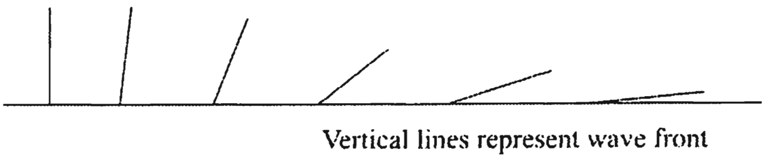


Figure 4.9 Tilt of a wave front as it propagates over the ground.

front tilts. So the surface wave dies more quickly as the frequency increases. Conversely at very low frequencies, the tilt angle can equal the curvature of the Earth and the surface wave will travel for very long distances.

4.8 Wave Diffraction

Physical diffraction is caused by a wave impinging on the edge of an object. Some energy in the wave appears to be bent by the edge leaving in a direction different from the original direction. Some of the energy in the wave is changed in its propagation direction. At VHF and above, this diffraction accounts for why it is possible to receive signals even though there are significant obstacles between the transmitter and receiver.

When a radio wave encounters an obstacle, a physical phenomenon called Huygen's principle explains the wave diffraction that occurs. This principle says that each source on a wave front generates secondary wave fronts called wavelets, and a new wave front is built from the vector sum of these wavelets. The amplitude of the wavelets varies as $(1 + \cos \alpha)$ where α is the angle between the wavelet direction and the direction of propagation. Thus, the wavelet with the maximum amplitude is in the direction of propagation and there is zero amplitude in the reverse direction. In schematic form, the waves are generated as shown in Figure 4.10. This is illustrated by points A and B in Figure 4.10.

Thus the obstacle does not totally block the signal behind it. It is assumed here that the obstacle has small enough thickness that any impacts at the end of the obstruction are insignificant. In the far field from the obstruction, the E-field at the bottom in the figure has flipped 180°. At the receiver antenna this accounts for the ability, in some cases, to receive a signal behind the obstruction.

Perhaps the most visible demonstration of Huygen's principle is at a small island in the ocean. Even though the waves at that point are predominantly in a single direction as prescribed by the wind or other natural factors, the waves lapping upon the beaches are as if the overall waves at that point are (almost) perpendicular to the beach, irrespective of the side of the island. When the obstacle is located such that the path-length difference between the direct path and the

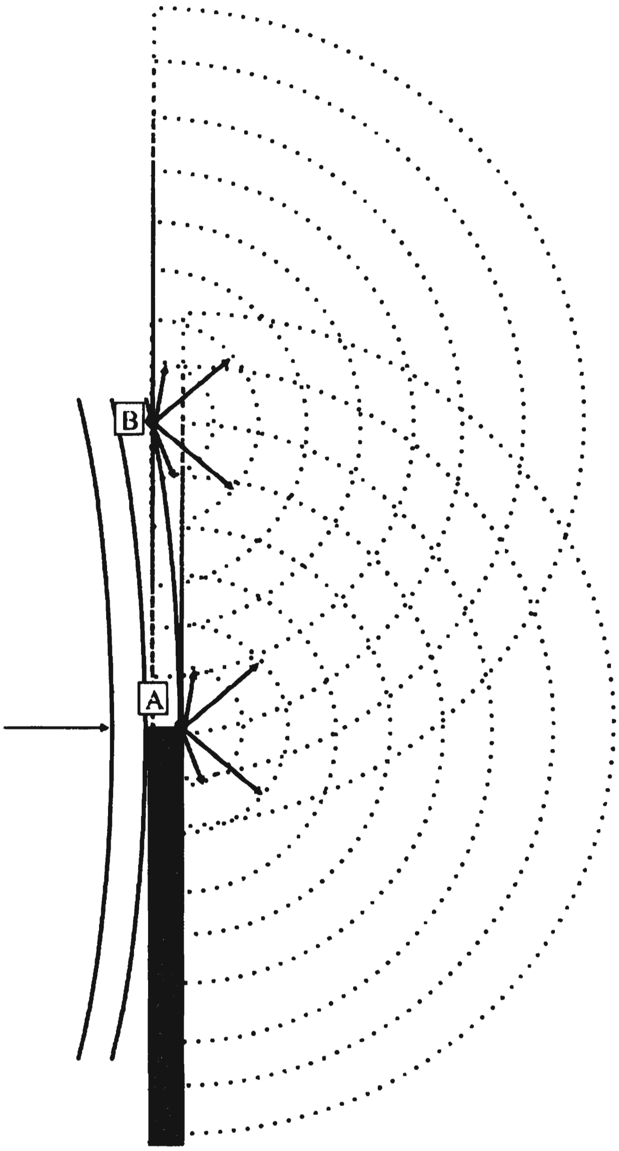


Figure 4.10 Diffraction at an obstacle produces signal energy behind the obstacle due to the Huygen's principal. The two points shown, A and B, effectively reradiate a wave as if they were the original source. This causes EM waves to appear behind the obstacle.

length of the path made up of the path from the transmitter to the obstacle and the path length from the obstacle to the receiver is a multiple of λ/2, then the diffraction from the obstacle will tend to cancel the direct path signal at the receiver. The reason for this is that the portion of the signal not blocked by the obstacle will generate signal vectors behind the obstacle but π radians out of phase with the direct path signal. The situation depicted in Figure 4.10, and with the assumption that the thickness of the obstacle has insignificant impact, is known as knife-edge diffraction, as previously mentioned. When there is a finite thickness, the attenuation at the edge can be substantially higher than at a knife edge. Diffraction effects also occur behind an obstruction such as a mountain.

4.9 Reflected Waves

The ground and other large surfaces (large relative to the wavelength of the signal) can reflect EM waves. Reflected waves can arrive at a receiver antenna out of phase with the direct wave, and, in the case where they are 180° out of phase, can cause considerable fading, depending on the magnitude of the reflected wave compared to the direct wave. These reflections can occur in many ways, one of which is off the ground when the transmitter and/or receiver are close to the Earth's surface. Another form of reflection is off nearby metallic objects, such as cyclone fences close to the transmitter and/or receiver. The ghosts that appear on a TV set (when connected to TV antenna as opposed to a cable or satellite antenna) are a manifestation of such reflections. In that case the reflected wave is received only slightly later than the direct wave, causing the picture to be delayed slightly. The geometrical shapes generated by setting the path difference equal to a constant are ellipsoids (three-dimensional ellipse) with the two antennas at the foci. When the path difference is set equal to $k\lambda/2$ for some integer k, the *Fresnel zones* are generated. For $k = 1$ it is the first Fresnel zone. These notions are depicted in

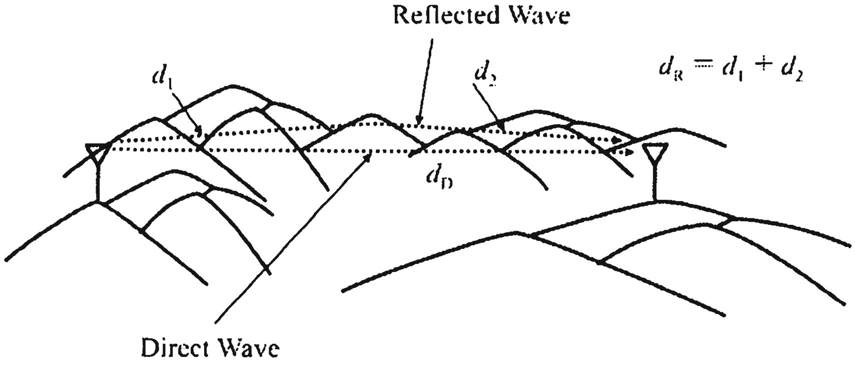


Figure 4.11 An EM wave with a significant reflected component at the receiver—the reflection off a distant mountain in this case.

Figure 4.11. Let $\delta_r = d_R - d_D$. Then the Fresnel zones are defined by

$$\delta_r = k\frac{\lambda}{2} \qquad k = 1, 2, \ldots \tag{4.34}$$

In many cases, it is only necessary to consider the first Fresnel zone.

Signal reflections off objects also affect the signal level received. At low reflection angles of incidence (the case for reflections off the Earth with large path distances and the case of importance here), reflections off the ground impart a π radian phase shift in the reflected signal for both vertical and horizontal polarizations. A rule of thumb useful for determining at what distance such reflections cause propagation changes from $n = 2$ to $n > 2$ in (4.27) is given by

$$d_t = \frac{4h_T h_R}{\lambda} \tag{4.35}$$

where h_T is the transmit antenna height, h_R is the receiver antenna height, and λ is the wavelength, all in consistent units. This corresponds to the distance where the first Fresnel zone first touches a point of reflection. Beyond this distance the signal is grazing the Earth at the reflection point and a constant π radians is destructively added to the phase of the received signal. The amplitudes of the respective signals determine the amount of the destruction, so therefore the characteristics of the reflection point (smooth or rough Earth and trees) are important.

Reflection of radio waves off the surface of the Earth can be analyzed with the aid of Figure 4.12. For simplicity it is assumed that the Earth is flat in the region between the transmitter and receiver. The power from the receiver antenna due to both the direct wave and the reflected wave is given by

$$P_{\text{total}} = P_{\text{direct}} \left| 1 + \rho e^{-j\phi} \right|^2 \tag{4.36}$$

where ρ is the reflection coefficient at the point of reflection and the phase difference between the direct wave and reflected wave at the receiver, ϕ, is given by

$$\phi = \omega\frac{\delta_r}{c} = 2\pi f\frac{\delta_r}{c} = 2\pi\frac{c}{\lambda}\frac{\delta_r}{c} \tag{4.37}$$

where δ_r is the path length difference between the two waves. Thus

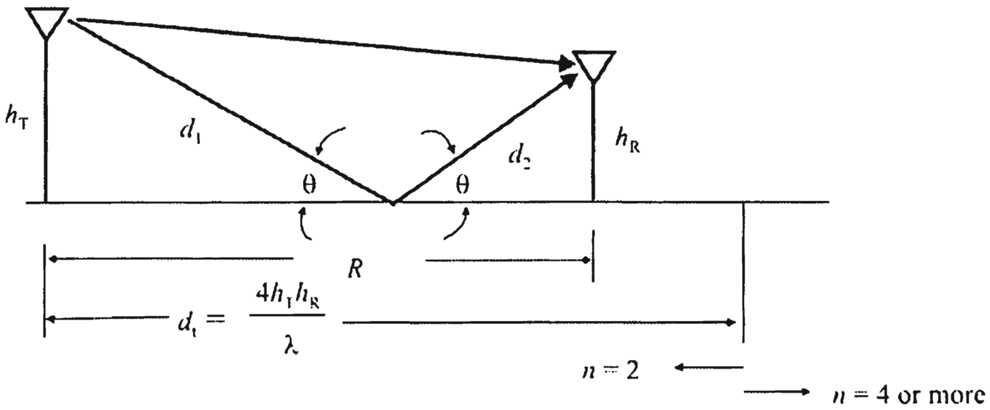


Figure 4.12 Distance d_1 is the distance where the wave first touches flat Earth for the first time. The first Fresnel zone is defined as that region around the direct path where the wave traveling over d_1 and d_2 is exactly 180° out of phase with the wave on the direct path. Beyond distance d_t, as shown, the attenuation exponent n changes from 2 to 4 or more.

$$\phi = 2\pi \frac{\delta_r}{\lambda} \tag{4.38}$$

If $r >> h_T$ or h_R then $\theta \approx 0$, $\rho \approx -1$, $\delta_r = 2h_T h_R / r$, and

$$\phi \approx \frac{2\pi}{\lambda} \frac{2h_T h_R}{r} \tag{4.39}$$

Now

$$\begin{aligned}
\left|1 + \rho e^{-j\phi}\right|^2 &= \left|1 - (\cos\phi - j\sin\phi)\right|^2 \\
&= (1 - \cos\phi)^2 + \sin^2\phi \\
&= \left(1 - 2\cos\phi + \cos^2\phi + \sin^2\phi\right) \\
&= (2 - 2\cos\phi) \\
&= 2 - 2\cos\left(\frac{2\pi}{\lambda} \frac{2h_T h_R}{r}\right)
\end{aligned}$$

so

$$\frac{P_{\text{total}}}{P_{\text{direct}}} = 2 - 2\cos\left(\frac{2\pi}{\lambda} \frac{2h_T h_R}{r}\right) \tag{4.40}$$

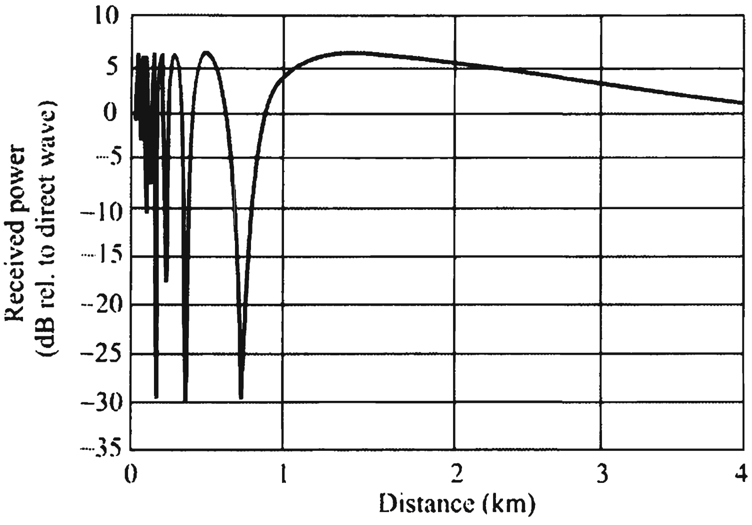


Figure 4.13 Power received due to the interaction of the direct wave and the ground reflected wave.

Equation (4.40) is plotted in Figure 4.13 in decibels relative to the direct wave for h_T = 30m and h_R = 2m at 1,850 MHz. When r is small the total power oscillates dramatically, but for larger ranges between the transmitter and receiver the power decreases at a rate proportional to $1/r^4$. The reason for this is as follows. For small ϕ, $\cos\phi \approx 1 - \phi^2/2$ so

$$\begin{aligned} \left|1+\rho e^{-j\phi}\right|^2 &= 2\left(1-1+\frac{\phi^2}{2}\right) \\ &= \phi^2 \\ &= \left(\frac{2\pi}{\lambda}\frac{2h_T h_R}{r}\right)^2 \end{aligned} \tag{4.41}$$

therefore

$$\left|1+\rho e^{-j\phi}\right|^2 \approx 16\left(\frac{\pi}{\lambda}\frac{h_T h_R}{r}\right)^2 \tag{4.42}$$

and

$$P_{\text{total}} = \frac{P_T}{4\pi r^2}\frac{\lambda^2}{4\pi}16\left(\frac{\pi}{\lambda}\frac{h_T h_R}{r}\right)^2$$

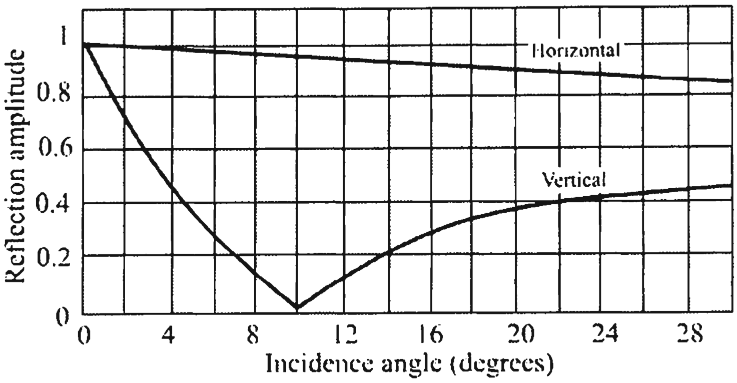


Figure 4.14 Amplitude characteristics of a reflected wave.

$$= \frac{P_{\mathrm{T}}}{r^4}\left(h_{\mathrm{T}}h_{\mathrm{R}}\right)^2 \tag{4.43}$$

for isotropic antennas. Note that the total power is independent of frequency. Furthermore, the total power increases as the square of the antenna height and decreases as the fourth power of the range.

When the antenna gains are considered, both for the transmitter and receiver, then this expression becomes (with G_{TR} and G_{RT} the gain relative to isotropic in the direction of each other)

$$P_{\mathrm{total}} = \frac{G_{\mathrm{TR}}G_{\mathrm{RT}}P_{\mathrm{T}}}{r^4}\left(h_{\mathrm{T}}h_{\mathrm{R}}\right)^2 \tag{4.44}$$

This expression is often referred to as the *two-ray model.* Although in (4.43) the received power is independent of frequency, when the gains are included, the power becomes frequency-dependent because the antenna gains are frequency-dependent.

Using this expression for the received power at all ranges underestimates the power at close ranges. Beyond distance d_{t}, given in (4.35), it more accurately reflects the total power received, however. At close range the propagation loss increases with $n = 2$ and the antennas are close enough together that the antenna heights do not have an effect. The propagation is effectively free-space.

Reflection amplitude characteristics are shown in Figure 4.14 when the frequency is 100 MHz and the reflections are off the ground. Horizontally polarized waves undergo substantially less amplitude attenuation than vertically polarized signals. Reflection phase characteristics are shown in Figure 4.15. A horizontally polarized wave always undergoes a 180° phase shift, while a vertically polarized one only imparts a significant phase shift for incidence angles less than 12° or so. The region of primary interest for ground-based

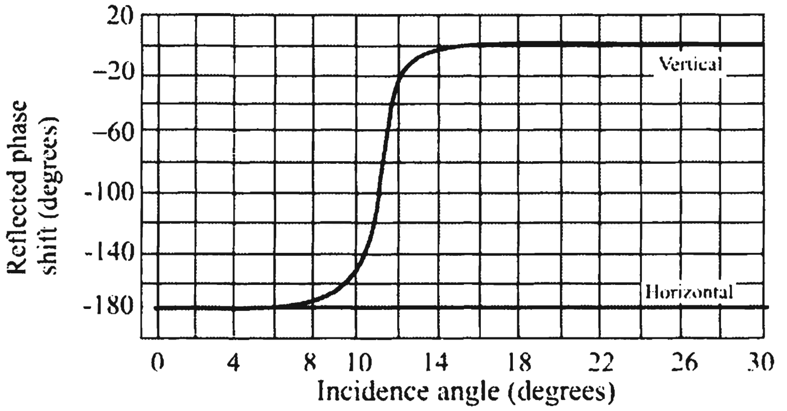


Figure 4.15 Phase shift imparted on a reflected wave.

communication EW system design is for low-incidence angles—typically less than 10°, while for airborne systems it can be substantially larger than this.

4.10 Ducting

There is a phenomena called *ducting*, wherein VHF and above signals can travel considerably farther than RLOS. Essentially the signals are reflected off regions in the troposphere where the refractive index decreases rapidly, forming a duct through which these signals will travel. This is shown in Figure 4.16. Two tropospheric layers or one layer and the Earth's surface can form such ducts.

In EW applications, ducting can be useful for ES from ranges and sites that otherwise would be too far away. On the other hand, determining the geolocation of such signals is difficult. This is because triangulation (defined in Section 11.3) usually will not work since ducts probably will not exist between the transmitter and two receiving sites simultaneously, a requirement for triangulation to function properly.

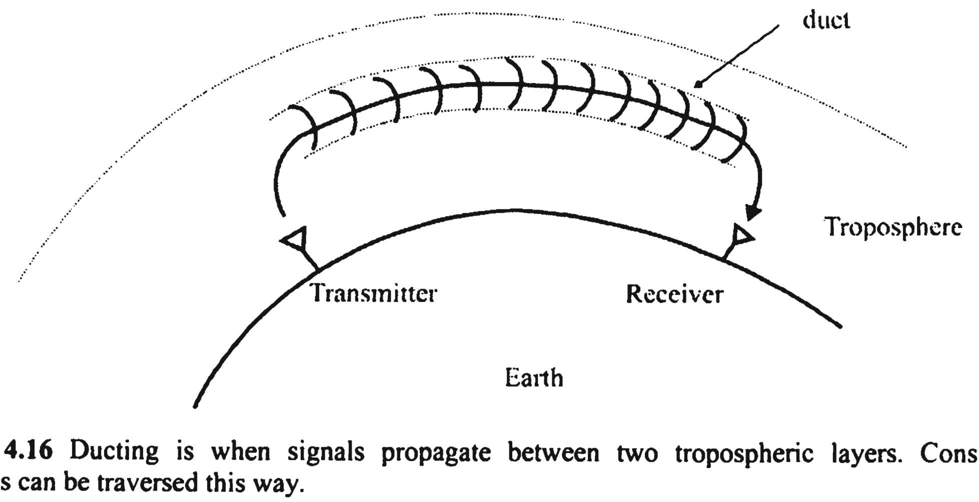


Figure 4.16 Ducting is when signals propagate between two tropospheric layers. Considerable distances can be traversed this way.

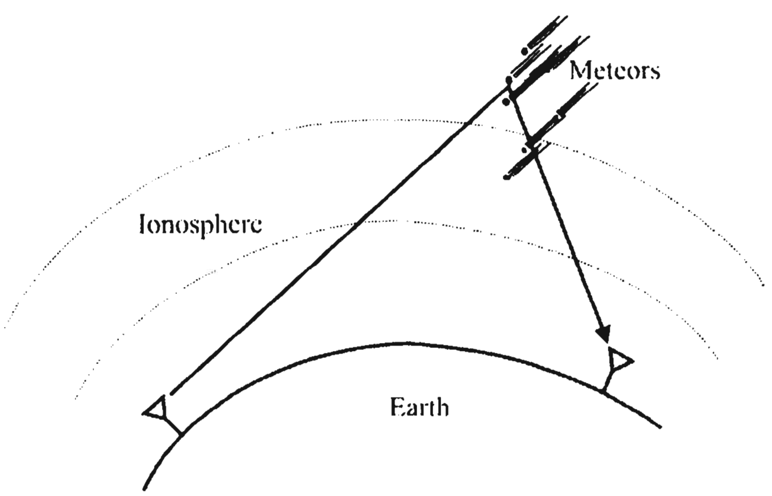


Figure 4.17 Meteor burst propagation reflects the EM wave off meteors. While considerable distance can be covered this way, data rates are quite low.

4.11 Meteor Burst

This form of communication relies on the thousands of meteors that enter the Earth's atmosphere each day [10]. It also can be classified as a reflected mode. At frequencies around 30–50 MHz, the ionized tails of these meteors will reflect radio waves. The phenomenon is illustrated in Figure 4.17.

Signals need to have considerable redundancy in this scheme, and therefore it is only reliable for low data rate communication. Communication is limited to short bursts and the medium only supports digital communications in the low data rate range—up to 600 bps is typical. It is frequently used for relaying remote sensor data where messages are infrequently sent, and each such message consists of only a few hundred bits maximum. Frequently *acknowledgments* (ACKs) and *negative acknowledgments* (NAKs) are used to indicate that a message was received properly. An ACK is sent if the message was received properly and a NAK is replied if not. If not, the message is repeated until it gets through, which means that one or more meteors was in the right place to reflect the signal. Of course, these ACKs and NAKs rely on the same propagation path so they too must be kept short and infrequent since the meteor could very likely no longer be there when the reply needs to be sent.

4.12 Scattering

Scatter-wave propagation is caused by nonhomogeneous refractive indices in the troposphere caused by irregular ionization, or by rain. It is also caused by nonhomogeneous refractive indexes on the surface of the Earth. Objects that are smaller than a wavelength will cause scattering for EM waves as well. Typical of

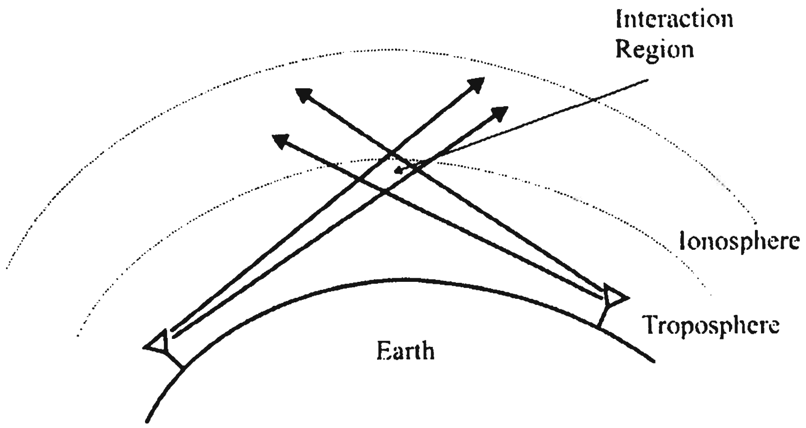


Figure 4.18 Tropospheric scatter propagation relies on the interaction to the two antenna beams at common areas within the troposphere. This type of propagation is possible in the 4–5 GHz range.

this latter category would be street signs and telephone posts scattering UHF signals, as typified by mobile phones in the 900 MHz range (λ = 0.3 meters) or *personal communication systems* (PCSs) in the 1,800 MHz range (λ = 0.15 meters). At frequencies around 4–5 GHz, there is a reliable propagation phenomena known as *tropospheric scattering*. A radio wave with properly oriented transmitting and receiving antennas can communicate over long distances. The configuration is shown in Figure 4.18. The transmitted signal is bent in the region where the antenna patterns overlap and the radio wave is essentially reflected from within this region.

4.13 Signal Fading

This section introduces signal fading phenomena; a good deal more will be said later in Chapter 13. Understanding how and why signals fade are important for understanding the effects on EW systems. Fading is caused by the interaction of several different propagation paths between the transmitter and receiver.

Generally motion of some form is required in order to experience time-dependent fading. Therefore, mobile transceivers as manifest in military C2 situations experience fading phenomena. Cellular phone systems and low-Earth orbit satellite communication systems also exhibit fading. It need not be the transceiver moving, however, that causes the fading. Large moving structures such as trucks and ships moving in the vicinity of the communication system can cause fading.

The basic characteristics of signal fading are presented first. The physics involved are discussed. That is followed by presentation of the more popular probabilistic mathematical models used in the system analysis for computing communication reliability.

The discussions in this section are largely focused on cellular phone systems and personal communication systems, although fading occurs independent of the frequency band and signal type. Cellular systems provide good examples, however, because many readers are familiar with them. In addition, the frequency is high enough that the effects of the sources of fading can be readily envisioned.[1] Lastly, cellular phone systems form a significant target set for urban warfare typical of the 21st century.

Fading considerations are important to understand for cellular systems for several reasons. Most mobile communication systems are used in and around centers of population. In such constructs, especially in large cities, the reflections and blockage caused by building structures are a principal cause of fading. The *base station* (BS) antennae are located on top of tall buildings or towers and they radiate at the maximum allowed power. On the other hand, the mobile antenna or *mobile station* (MS) is well below the surrounding buildings and the power levels are carefully controlled, at least in some types of systems.

Wireless communication fading phenomena are mainly due to scattering of electromagnetic waves from surfaces or diffraction over and around buildings. The design goal is to make the received power adequate to overcome background noise over each link, while minimizing interference to other more distant links operating at the same frequency.

4.13.1 Signal Model

We model a narrowband propagation channel by considering a sinusoidal transmitted carrier

$$s(t) = A\cos\omega_c t \tag{4.45}$$

This signal received over a Ricean multipath channel can be expressed as

$$r(t) = C\cos\omega_c t + \sum_{n=1}^{N} r_n \cos(\omega_c t + \theta_n) \tag{4.46}$$

where C is the amplitude of the line-of-sight component, r_n is the amplitude of the nth reflected wave, θ_n is the phase of the nth reflected wave, and $n = 1, \ldots, N$

[1] Such effects are observable over single or a few wavelengths which are on the order of a meter in the cellular bands.

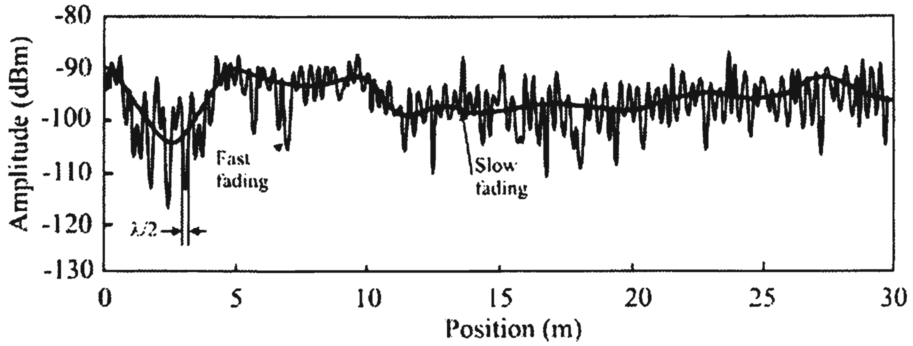


Figure 4.19 Fast and slow fading.

identify the reflected, scattered waves. Note that Rayleigh fading corresponds to $C = 0$.

4.13.2 Types of Fading

Multiple propagation paths create fading effects at the receiver. Depending on the characteristics of the paths, fast as well as slow fading may be experienced by a received narrowband signal. Many such *multipath signals* can exist at any given time. Fast and slow fading are illustrated in Figure 4.19.

The effects of reflections, blocking, and scattering are exemplified in Figure 4.20. As the MS moves along the street, rapid variations of the signal are found to occur over distances of about one-half the wavelength $\lambda = c / f$. Here c is the speed of light and at $f = 910$ MHz for example (λ = 32 cm). Over a distance of a few meters the signal can vary by 30 dB or more. Over distances as small as $\lambda / 2$, the signal may vary by 20 dB. *Fast*[2] *fading* occurs when there are rapid fluctuations of the signal over small areas. The multiple rays set up an interference pattern in space through which the MS moves. When the signals arrive from multiple directions in the horizontal plane, fast fading will be experienced for all directions of motion.

Reflections off buildings or other large structures in the vicinity of the transmitter or receiver will cause a change in the median value about which the rapid fluctuations take place. This signal variation, which is influenced by dimensions on the order of the building's size, is known as *shadow fading*, *slow fading*, or *lognormal fading*.

The channel exhibits *flat fading* if it has constant gain and linear phase response over a bandwidth which is greater than the bandwidth of the transmitted signal. Flat fading occurs when the bandwidth of the transmitted signal B is

[2] Fast and slow refers to time durations relative to the symbol time.

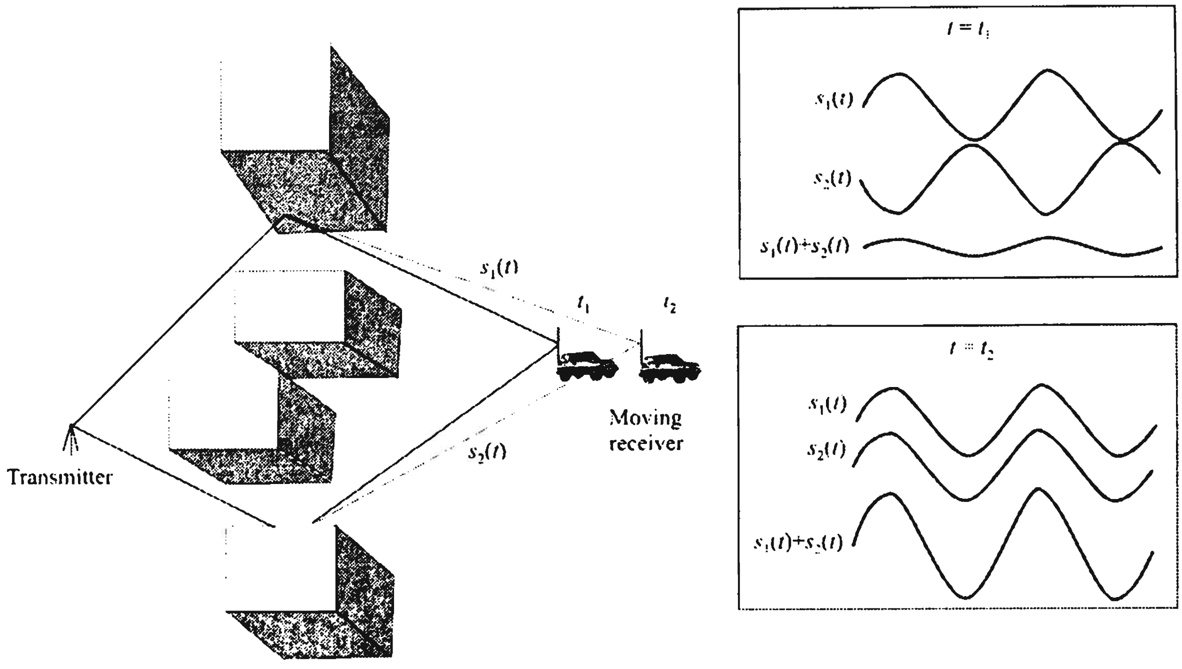


Figure 4.20 Multipath channels created by reflections off of large objects. If one or more reflecting object or the transmitter or receiver is moving (the receiver in this case) the fading is nonstationary.

substantially smaller than the coherence bandwidth[3] of the channel $B << B_c$. The effect of flat fading is to reduce the SNR.

4.13.2.1 Fast Fading

Fast fading occurs if the channel impulse response changes rapidly within the symbol duration—it is defined as when the coherence time[4] of the channel T_D is significantly smaller than the symbol period of the transmitted signal $T_D << T$. This causes frequency dispersion or time selective fading due to Doppler spreading where the channel reacts to different frequencies differently. Fast fading is illustrated in Figure 4.19. It is due to reflections off local objects and the motion of the MS relative to those objects. The receive signal is the sum of a number of signals reflected from local surfaces, and these signals sum at the receiver in a constructive or destructive manner caused by relative phase shifts. This phase relationship depends on the speed of motion, frequency of transmission, and relative path lengths.

[3] The coherence bandwidth of a channel, B_c, is approximately the inverse of RMS delay spread of a multipath channel. It can be thought of as the range of frequencies over which the channel does not significantly affect the spectral amplitude and the phase is linear.

[4] The *coherence time* is the expected time duration over which the channel's response is essentially invariant.

Even though Figure 4.20 shows a reflector moving to cause the fading effects, any or all of the constituent parts may be moving and the fading effects will be experienced. The transmitter could move behind buildings or move to a point where the reflection off a structure is totally different. Likewise, although buildings don't normally move, a semitrailer truck does and they are the size of small buildings.

In addition, as mentioned above, fading effects are present in every communication system to some extent. For example, the motion of the ionospheric layers and the subsequent changes in their reflection (refraction actually) characteristics will cause HF signals to experience very deep fades—to the point of precluding communications at all.

Likewise for meteor burst communication systems where signals are reflected from meteors to facilitate communications. The signal level at the receiver depends on the number and size of the meteors providing reflection at the moment. Even though there are a large number of meteors that enter the Earth's atmosphere every day, the number providing reflection for a given link varies considerably from moment to moment, thus producing varying multipath paths and the resulting fading effects.

4.13.2.2 Slow Fading

Slow fading is manifest in a slowly varying (relative to the fast fading described above) median value of the received signal. It is caused by shadowing of the receiver as it moves behind buildings, mountains, hills, and other large objects. The average fading within individual small areas also varies from one small area to the next in an apparently random manner. While there is no adequate mathematical model available for slow fading, it does frequently follow a log-normal distribution so is sometimes called log-normal fading.

To separate fast fading from slow fading, the envelope or magnitude of the MS receiver signal is averaged over a distance significantly larger than a wavelength (e.g., 10m). Alternatively, a sliding window can be used.

4.13.3 PDF of Amplitude, Phase, and Power for Rayleigh Fading

A signal comprised of several multipath components but without a significant deterministic component is illustrated in Figure 4.21. Such a signal carries the appellation, a *Rayleigh fading signal.* The received signal, given by the dark vector, consists of $N = 5$ reflected waves (lighter). The resulting signal amplitude ρ consists of an in-phase component I and a quadrature component Q. If the receiver moves, the relative phases of the reflected waves change over time, so ρ, I, and Q become functions of time t. The Pythagorean Theorem says that the amplitude is given as

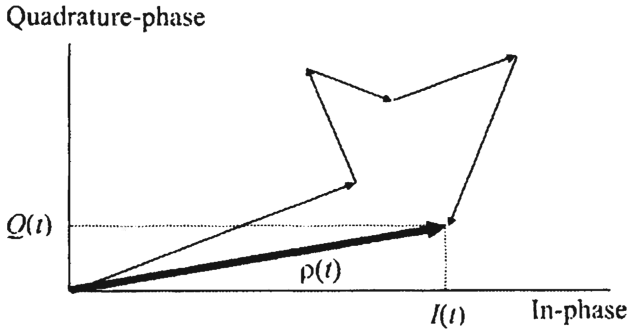


Figure 4.21 Rayleigh fading signal where the signal is comprised of several random components and no dominant, deterministic component.

$$\rho(t) = \sqrt{I^2(t) + Q^2(t)} \tag{4.47}$$

When the number of received waves N becomes very large and all are *independent and identically distributed* (i.i.d.), from the central limit theorem we know that at $t = t_0$, $I(t_0)$ and $Q(t_0)$ are zero-mean, Gaussian random variables, each with variance σ^4. The received signal is given by

$$r(t) = \text{Re}\{\rho(t)e^{j\omega_c t}\} = \rho(t)\cos[\omega_c t + \theta(t)] \tag{4.48}$$

with $\rho(t)$ given by (4.47), and a uniform phase $\theta(t)$ between 0 and 2π so that

$$f_\Theta(\theta) = \begin{cases} \dfrac{1}{2\pi}, & 0 \le \theta \le 2\pi \\ 0, & \text{elsewhere} \end{cases} \tag{4.49}$$

The pdf of the Rayleigh amplitude $\rho(t)$ is derived as follows. Let $z = x + jy$ with x and y i.i.d., Gaussian with zero mean and variance σ^2. Thus the joint pdf of x and y is given by

$$f_W(w) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) \tag{4.50}$$

Note that $|z| = \sqrt{x^2 + y^2}$ so that when $x = I(t)$ and $y = Q(t)$ then $\rho = |z|$. Because of the Gaussian nature of x and y, the probability that $(x^2+y^2)/2 < w$, with power value w is

$$F_W(w) = \frac{1}{\sqrt{2\pi\sigma^2}} \iint\limits_{x^2+y^2<2w} \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) dxdy \tag{4.51}$$

Converting to polar coordinates using

$$dxdy = \rho d\rho d\theta \tag{4.52}$$

we get

$$F_W(w) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_0^{2\pi} d\theta \int_0^{\sqrt{2w}} \rho \exp\left(-\frac{\rho^2}{2\sigma^2}\right) d\rho \tag{4.53}$$

Thus,

$$F_{\mathrm{W}}(w) = 1 - \exp\left(-\frac{w}{\sigma^2}\right) \tag{4.54}$$

This is the (exponential) distribution of received power w, popularly known as the *Rayleigh distribution function*. The pdf is found by taking the derivative of (4.54) with respect to w and we get

$$f_W(w) = \frac{1}{\sigma^2} \exp\left(-\frac{w}{\sigma^2}\right) \tag{4.55}$$

The instantaneous power w thus has the above exponential pdf.

Conversion between the probability density of amplitude, ρ, and that of power, w, requires that

$$|f_P(\rho)| = |f_W(w)| \tag{4.56}$$

and $w = \rho^2/2$, so $dw = \rho d\rho$. Making that transformation we get

$$f_W(w) = f_P(\sqrt{2\rho}) \left|\frac{d\rho}{dp}\right| = \frac{1}{\sigma^2} \exp\left(-\frac{p}{\sigma^2}\right) \tag{4.57}$$

The pdf of the amplitude is

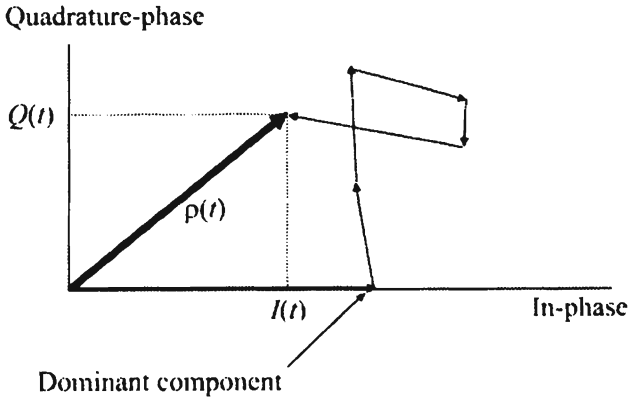


Figure 4.22 Ricean fading includes a dominant component, shown here as coincident with the in-phase random components.

$$f_P(\rho) = \frac{\rho}{\sigma^2}\exp\left(-\frac{\rho^2}{2\sigma^2}\right) \tag{4.58}$$

for $\rho > 0$, which is known as the *Rayleigh probability density function*. Much more will be said in the sequel about these functions.

4.13.4 PDF of Amplitude and Phase for Ricean Fading

The model for Ricean fading is similar to that for Rayleigh fading, except that in Ricean fading a strong dominant component is present as illustrated in Figure 4.22. This dominant component can be, for example, the line-of-sight signal or it could the phasor sum of two or more dominant signals, e.g., the line-of-sight, plus a ground reflection. This combined signal is then normally treated as a deterministic process. The dominant wave can also be subject to shadow attenuation, where the receiver or transmitter moves behind a significant obstacle. This is a popular model for satellite channels. Besides the dominant component, the mobile antenna receives a large number of reflected and scattered waves as illustrated in Figure 4.23.

Ricean fading is a popular model for vehicle-to-vehicle communications, common for military targets. It is also a good model for very small cellular systems (microcellular channels) where the MS are close enough to the base station to receive the dominant component. It is also a good model for indoor propagation channels and satellite channels, where, in both cases, there is typically a strong direct-wave component.

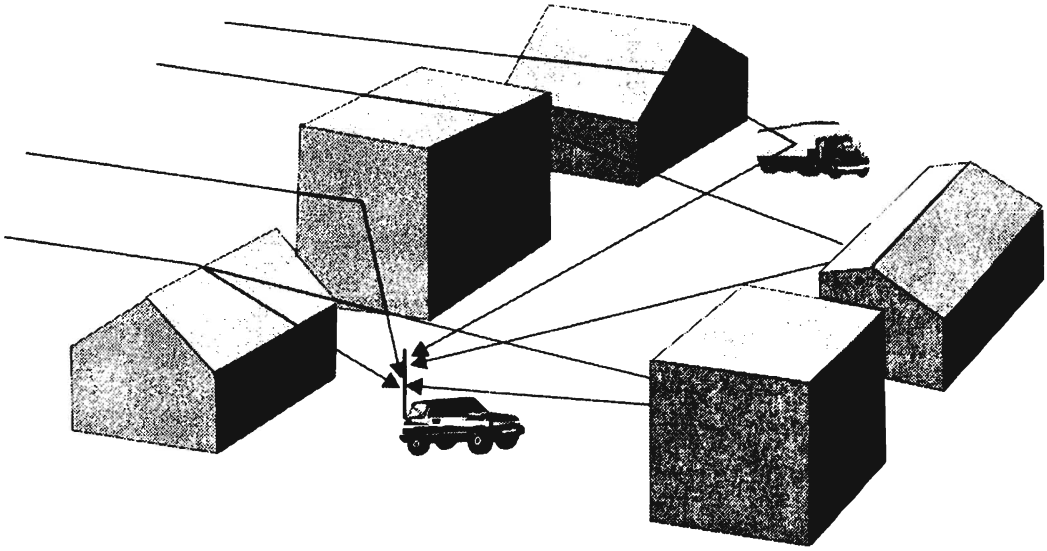


Figure 4.23 Reflections—in this case there is no direct signal. Every component is either diffracted or reflected (or both).

The dominant component illustrated in Figure 4.22 is shown inphase. Similar to the case of Rayleigh fading, the inphase and quadrature phase components of the received multipath signals are i.i.d. jointly Gaussian random variables. However, in Ricean fading the mean value of (at least) one component is nonzero due to the deterministic strong wave.

4.13.4.1 PDF of Ricean Signal Amplitude

It is frequently proposed to approximate the Ricean amplitude pdf by the model for Nakagami fading; however, the behavior of Nakagami and Ricean fading in deep fades is fundamentally different. Approximations that focus on the behavior near the mean value, where the two models are similar, will be in error by orders of magnitude in predicting the probability of deep fades.

The pdf of the signal amplitude $\rho(t)=\sqrt{I^2(t)+Q^2(t)}$ is determined by measuring the random processes $I(t)$ and $Q(t)$ at one particular instant t_0. Just as for the Rayleigh model, if the number of scattered waves is sufficiently large, and is i.i.d., the central limit theorem says that $I(t_0)$ and $Q(t_0)$ are Gaussian, but, due to the deterministic dominant term, one or both of them is no longer zero mean. By converting to polar coordinates, the joint amplitude-phase pdf is determined to be [11]

$$f_{P,\Theta}(\rho,\theta)=\frac{\rho}{2\pi\sigma^2}\exp\left(-\frac{\rho^2-2C\rho\cos\theta+C^2}{2\sigma^2}\right) \tag{4.59}$$

Here, σ^2 is the local-mean scattered power[5] and $C^2/2$ is the power of the dominant component. The pdf of the amplitude is found by integrating (4.59) over all possible phase values

$$\begin{aligned}f_P(\rho)&=\int_{-\pi}^{\pi}f_{P,\Theta}(\rho,\theta)d\theta\\&=\frac{\rho}{\sigma^2}\exp\left(-\frac{\rho^2+C^2}{2\sigma^2}\right)I_0\left(\frac{C\rho}{\sigma^2}\right)\end{aligned} \tag{4.60}$$

where $I_0(.)$ is the modified Bessel function of the first kind and zero order, defined as

$$I_0(x)\triangleq\frac{1}{2\pi}\int_{-\pi}^{\pi}e^{x\cos\theta}d\theta \tag{4.61}$$

Equation (4.61) has no known closed form solution and must be computed numerically.

Ricean K-Factor

The *Ricean K-factor* is defined as the ratio of signal power in the dominant component to the local-mean scattered power. In the expression for the received signal, the power in the line-of-sight component is $C^2/2$. Thus[6]

[5] The *local-mean scattered power* is the average of the power in all of the multipath signals.

[6] In one model [14] the K-factor distribution was found to be log-normal, with the median function of season, antenna height, antenna beamwidth, and distance given by

$$K=F_sF_hF_bK_0d^{\gamma}u$$

F_s is the seasonal factor = 1 in summer and 4.5 in winter; F_h is the receiving antenna height factor = $(h/3)^{0.46}$, h in meters; F_b is the antenna beamwidth factor = $(b/17)^{-0.62}$, b in degrees; d is the distance in km; γ is the exponent = –0.5; K_o is the 1 km intercept = 10 dB; u is the zero-mean log-normal variate with a 8.0 dB standard deviation over the area.

$$K \triangleq \frac{C^2}{2\sigma^2} \tag{4.62}$$

The total local-mean power is the sum of the power in the LOS and the local-mean scattered power:

$$\bar{w} = \frac{1}{2}C^2 + \sigma^2 \tag{4.63}$$

The local-mean scattered power equals $\sigma^2 = \bar{w}/(K+1)$. The amplitude of the line-of-sight component is

$$C = \sqrt{2K\bar{w}/(K+1)} \tag{4.64}$$

In indoor channels with an unobstructed LOS between transmit and the receiver antenna the K-factor is between, say, 4 and 12 dB. Since there is no dominant component with Rayleigh fading, then $C = 0$ yielding $K = 0$ ($-\infty$ dB).

Expressed in terms of the local-mean power $\bar{w}$ and the Ricean K-factor, the pdf of the signal amplitude becomes

$$f_P(\rho) = (1+K)e^{-K}\frac{\rho}{\bar{w}}\exp\left(-\frac{1+K}{2\bar{w}}\rho^2\right)I_0\left(\sqrt{\frac{2K(1+K)}{\bar{w}}}\rho\right) \tag{4.65}$$

Let $p = \rho^2/2$, so that $dp = \rho\, d\rho$. Since $f_W(w)\, dw = f_P(\rho)\, d\rho$ then

$$f_W(w) = \frac{(1+K)e^{-K}}{\bar{w}}\exp\left(-\frac{1+K}{\bar{w}}w\right)I_0\left[\sqrt{4K(1+K)\frac{w}{\bar{w}}}\right] \tag{4.66}$$

The Ricean amplitude pdf is plotted in Figure 4.24 for typical values of K.

4.13.4.2 PDF of Ricean Phase

It is often assumed that the phase in Ricean fading is uniform on $[0, 2\pi)$. This is, in fact, an invalid assumption. The Ricean fading phase pdf is given by

$$f_\Theta(\theta) = \frac{1+2\sqrt{\pi}e^{4k\cos^2(\theta-\phi)}}{2\pi e^{4k}}\{1+\operatorname{erf}[2\sqrt{k}\cos(\theta-\phi)]\} \tag{4.67}$$

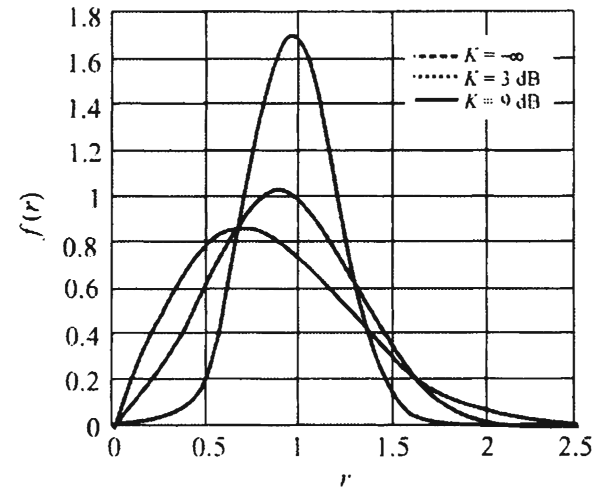


Figure 4.24 Ricean amplitude fading pdf.

where erf(.) is the error function,[7] $k \geq 0$ is the Rice fading parameter and ϕ is a phase term which depends on the ratio of the quadrature dominant component and in-phase dominant component. If they are identical in modulus then $\phi = \pi/4, 3\pi/4, 5\pi/4, 7\pi/4$. Figure 4.25 shows the shapes of the Ricean phase distributions for the various ranges of their parameters. Also shown in Figure 4.25 is the constant uniform $1/2\pi$ curve; clearly the constant assumption is not accurate.

4.13.5 PDF of Amplitude, Phase, and Power for Nakagami-*m* Fading

The Nakagami-*m* distribution is another frequently used model for the received signal amplitude in multipath propagation channels. The Nakagami fading model was initially proposed because it matched empirical results for short wave (HF) ionospheric propagation.

If the envelope is Nakagami distributed, the corresponding instantaneous power given by the square of the envelope is gamma distributed given by [12]

$$p_W(w) = \left(\frac{m}{\bar{w}}\right)^m \frac{w^{m-1}}{\Gamma(m)} \exp\left(-\frac{mw}{\bar{w}}\right), \qquad w \geq 0, m \geq \frac{1}{2} \qquad (4.68)$$

[7] The error function is given by

$$\text{erf}(z) \triangleq \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt$$

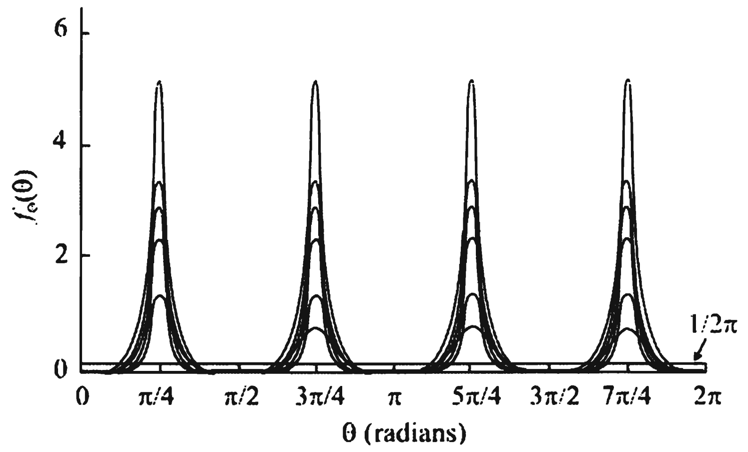


Figure 4.25 Ricean phase pdf.

The parameter m is called the *shape factor* of the Nakagami or the gamma distribution and $\bar{w} = \mathcal{E}\{w\}$ is the average signal power. The worst fading condition by this distribution is obtained for m = ½. The Rayleigh fading case can be fitted when m = 1 with an exponentially distributed instantaneous power. No fading occurs as $m \rightarrow \infty$. For $m > 1$, the variations in the signal strength reduce compared to Rayleigh fading. A few examples of this pdf are illustrated in Figure 4.26.

4.13.5.1 Nakagami-*m* Joint Envelope-Phase Distribution

It is frequently assumed that the Nakagami phase distribution is uniform over (0, 2π). It will be shown that this assumption is incorrect. The following derivation follows [12] closely.

Let R and Θ be random variates representing the envelope and phase, respectively, of the Nakagami-m signal. The corresponding joint pdf $p_{R,\Theta}(r, \theta)$ is given by

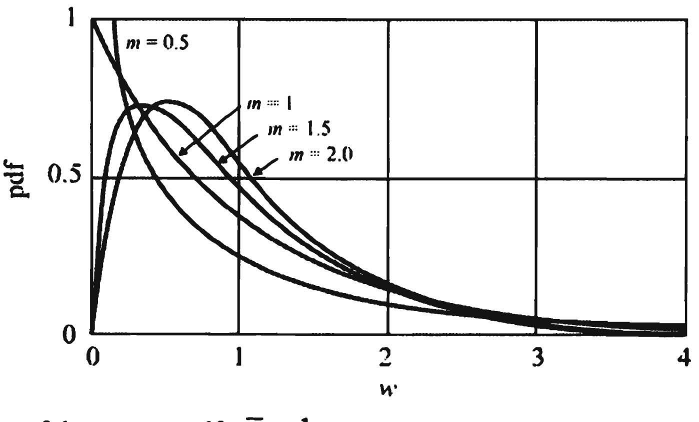


Figure 4.26 Examples of the gamma pdf, $\bar{w} = 1$.

$$f_{\mathrm{R},\Theta}(r,\theta)=\frac{m^m\left|\sin(2\theta)\right|^{m-1}r^{2m-1}}{2^{m-1}(2\sigma^2)^m\Gamma^2(m/2)} \tag{4.69}$$

where $2\sigma^2=\mathcal{E}\{r^2\}$, $\Gamma(.)$ is the Gamma function, and $m\geq 1/2$ is the fading figure. The envelope pdf $f_R(r)$ is determined by integrating (4.69) over all possible values of θ yielding

$$f_R(r)=\int_{-\pi}^{\pi}f_{R,\Theta}(r,\theta)d\theta=\frac{2m^m r^{2m-1}}{(2\sigma^2)^m\Gamma(m)}\exp\left(-\frac{mr^2}{2\sigma^2}\right) \tag{4.70}$$

The phase pdf $p_\Theta(\theta)$ is given by integrating (4.69) over all possible values of the envelope yielding

$$f_\Theta(\theta)=\int_{-\infty}^{\infty}f_{R,\Theta}(r,\theta)dr=\frac{\Gamma(m)\left|\sin(2\theta)\right|^{m-1}}{2^m\Gamma^2(m/2)} \tag{4.71}$$

Clearly

$$f_{R,\Theta}(r,\theta)=f_R(r)f_\Theta(\theta) \tag{4.72}$$

i.e., phase and envelope are independent random variates.

This phase distribution is illustrated in Figure 4.27. There are peaks in the pdf around $n\pi/4$ for $m\geq 1$, and n = 1, 3, 5, and 7; and at $n\pi/2$ for $m<1$ n = 0, 1, 2, and 3. Comparing Figure 4.27 with Figure 4.25 it is seen that Ricean fading and Nakagami fading have similar shapes and are coincident at the limiting cases of their parameters. Figure 4.28 shows the shapes of the Nakagami-*m* phase shapes in polar coordinates. The Nakagami pdf approximates the amplitude pdf of a signal received over several multipath channels after maximum ratio diversity combining (MRC).[8]

With *k*-branch MRC in Rayleigh-fading channels, the resulting signal exhibits a Nakagami pdf with $m=k$. MRC combining of *m*-Nakagami fading signals in *k*

[8] *Maximum ratio combining* (MRC) is a way of combining signals received over multipath channels to optimally receive transmitted signals. It is a special form of general diversity combining, by which multiple replicas of the same information-bearing signal received over different diversity branches are combined to maximize the instantaneous SNR at the combiner output. It is a popular technique in cellular communication systems.

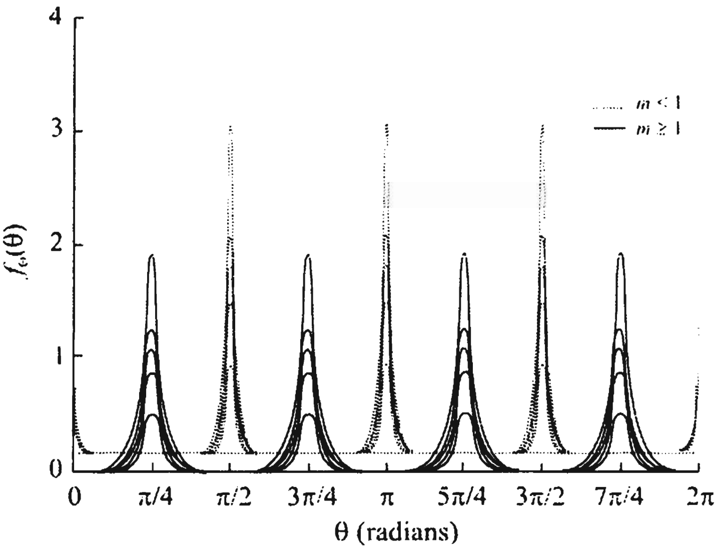


Figure 4.27 Nakagami phase pdf.

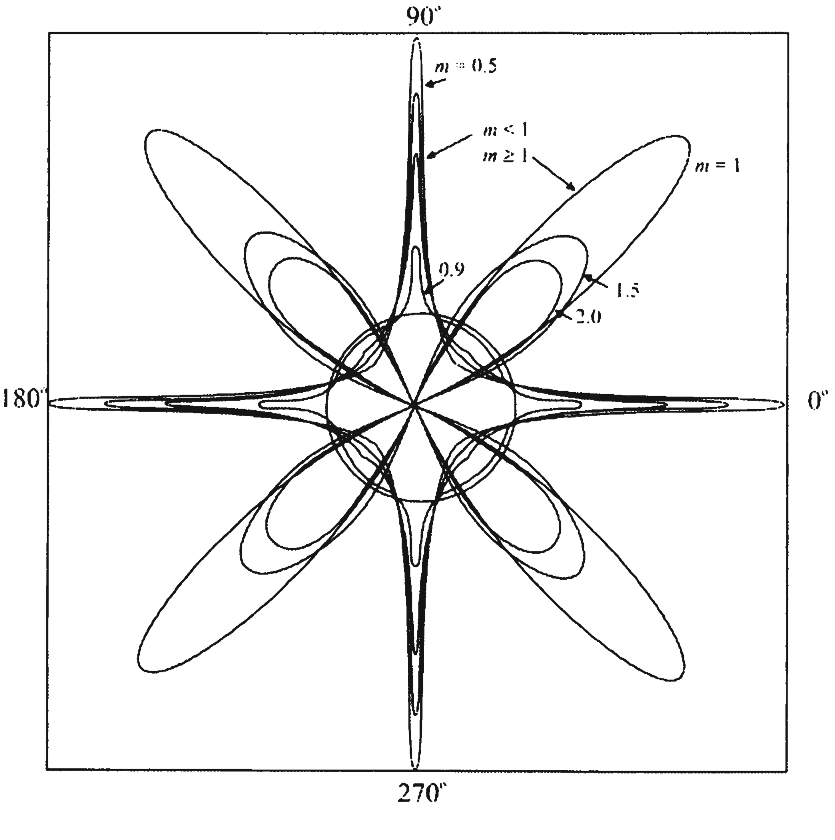


Figure 4.28 Nakagami phase pdf polar plot (viewed from above).

branches gives a Nakagami signal with shape factor *mk*. In addition, the sum of multiple i.i.d. Rayleigh-fading signals has a Nakagami distributed signal amplitude pdf. This is particularly relevant to modeling interference from multiple sources in an urban environment.

4.14 Characteristics of the Mobile VHF Channel

The previous results for propagation as applied to the mobile VHF channel are summarized in this section.

Command and control of mobile, tactical military forces frequently is accomplished with RF communications, primarily in the VHF and UHF ranges. This type of communication is subject to slow and fast fading and distortion due to delay spread. As pointed out earlier $P_{\text{den}} \propto 1/r^{n}$. The value of n depends on several factors, such as the ground conditions in the path between the transmitter and receiver. Illustrated at the top of Figure 4.29 is the characteristic path loss for $n = 2$, 3 and 4.

At any given distance from the transmitter, there will be a statistical distribution of the path loss, which therefore imparts the statistical distribution onto the amount of power received. A Gaussian distribution (see Appendix A) as shown often accurately describes this distribution. Objects blocking the direct communication path, called shadowing, typically cause large fades. The mean path loss at some distance R due to shadowing produces slow fades as shown in Figure 4.29. It causes (relatively) long-term variations in signal level at the receiver. This is typical of mobile communications where the receiver or transmitter moves behind large objects (e.g., a mountain). Small fades are caused by locally changing conditions such as changing multipath conditions.

On the other hand, fast fading also occurs in mobile communications. The amplitude of the received signal level is often approximated by a Rayleigh distribution, which is also shown in Figure 4.29, middle right panel. The receiver or transmitter moving behind small objects, causing the signal to be attenuated, produces such fading.

In direct and surface-wave propagation, fading can occur in mobile communication paths due to reflections off of large objects or changing diffraction or scattering situations. Even if the transmitter and receiver are not moving, objects moving in the environment surrounding them, such as vehicles or people, can change the propagation conditions. This fading can affect very narrow frequency bands, sometimes affecting only portions of the bandwidth of a signal, attenuating some portions of the frequency band much more than others. The longer-term attenuation characteristics of the ionosphere are attributable to the changing ionosphere, whereas fast fading occurs for other reasons.

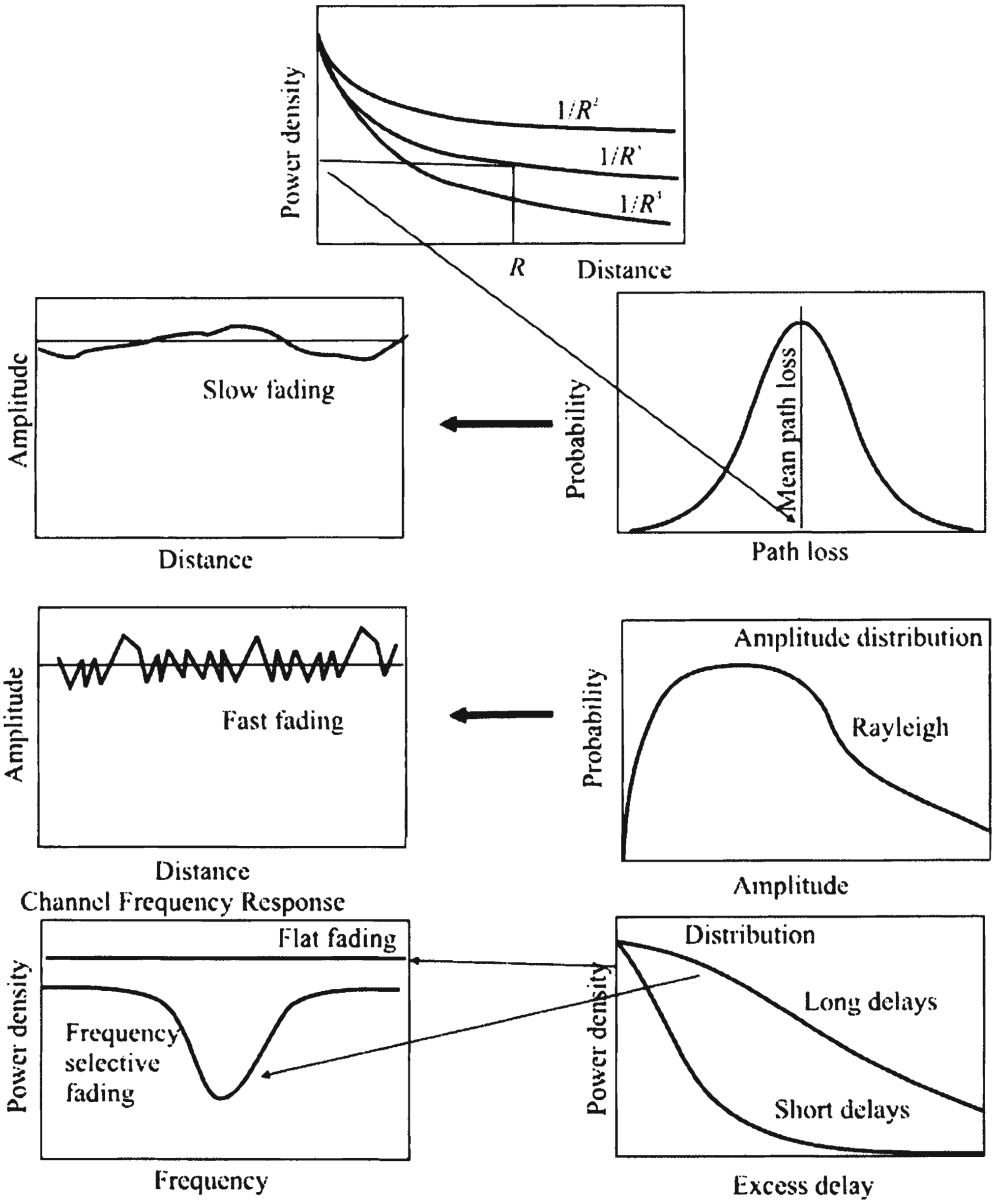


Figure 4.29 The mobile channel is characterized by slow and fast fading and stochastic propagation losses.

The paths taken by reflected waves compared to the direct wave, such as the situation shown in Figure 4.11, are different. These waves add as vectors at the receiver causing the reflected wave to cancel to some degree the direct wave. This path length, of course, varies as the transmitter and/or receiver moves, causing variable fading effects. Even though an EM wave has a well-defined polarization as it leaves the transmitting antenna, the effects of the environment change this polarization as the EM wave propagates. A wave reflected off the Earth or some other surface can change the direction of the polarization and can change its direction of propagation. Typically reflected or refracted waves arrive at a receiver with elliptic orientation. Furthermore, this elliptic orientation varies with time due to the changing propagation conditions. If the receiver antenna is oriented in a particular direction, which it normally is, this changing elliptic polarization will be received as fading phenomena. We will discuss fading and its impact on EW systems in considerably more depth in Chapter 13.

The RF channel can also cause time delays due to its impulse response. These time delays can be short or long, depending on the source of the delay. The delay is referred to *excess delay*, where excess is relative to the symbol rate. Long delays are caused by a frequency-selective impulse response. Short delays are due to a relatively flat channel frequency response. The frequency referred to in this case is relative to the bandwidth of the signal.

Reflected waves arrive at a receiver later than the direct or surface wave. This delay time is referred to as the *delay spread* and varies depending on the environmental conditions. In urban environments the delay spread is typically 3 μs, in suburban environments it is typically 0.5 μs while in rural terrain it can vary considerably, ranging from less than 0.2 μs to 12 μs or more. These values are means since the delay spread is a random variable in most circumstances. The standard deviations of these distributions can vary considerably depending on the specific situation.

This multipath interference tends to smear digital signals and cause one symbol to interfere with others called *intersymbol interference* (ISI). This interference can be quite severe and limits the maximum data rate that the communication channel can support. In an urban channel, for example, when the delay spread is 3 μs, digital signals at 333 kbps would have symbol n in the direct wave completely overlaid by the $n - 1$ symbol in the delayed signal. If the energy in the delayed signal is strong enough, reliable communication in that case would be impossible.

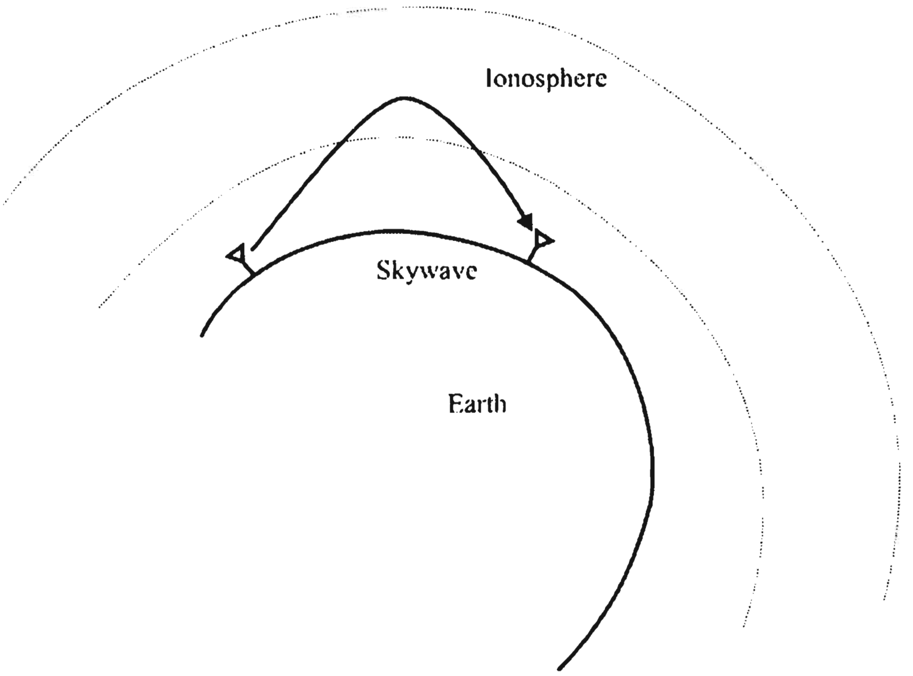


Figure 4.30 Long-range propagation is possible by signal refraction in the ionosphere.

4.15 Propagation via the Ionosphere

Propagation of communication signals via refraction and reflection from the ionosphere is one of the oldest, if not *the* oldest, forms of radio communication [13]. It is certainly the oldest commercial communication method. The signals that radiate quite a distance by refraction through the ionosphere are sometimes called *skywave* signals. These signals can frequently propagate for thousands of kilometers from the transmitter. Such propagation modes have been useful for communication with ships at sea and were virtually the only way to communicate with ships until communication satellites were developed [7, 14].

It is primarily the radiation from the Sun that generates the ions in the ionosphere. Therefore, there are differences in the propagation characteristics of the ionosphere depending on whether it is day or night. At dawn and dusk the ionosphere is turbulent due to the changing sun radiation situation. The density of these ions varies with altitude and time of day.

4.15.1 Refraction

Just as VHF and above signals are refracted by the troposphere as discussed earlier, high-frequency signals are refracted by the ionosphere. At frequencies below about 30 MHz, the ionosphere (altitude 50–500 km) can refract signals, as illustrated in Figure 4.30. Whether it does so depends on several factors. The

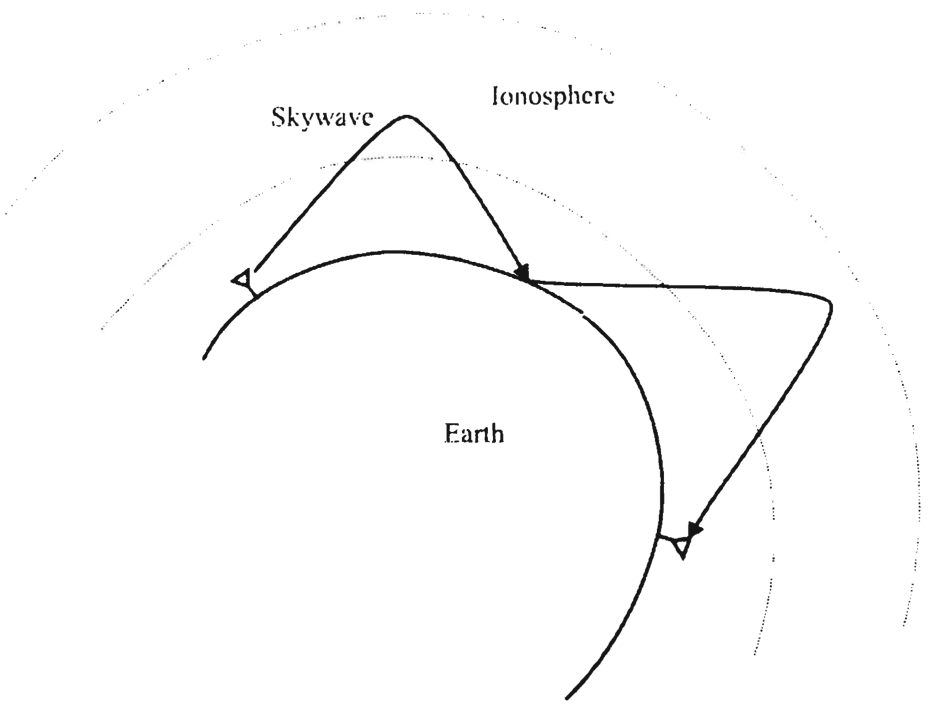


Figure 4.31 Multihop ionospheric refraction is a method to propagate HF signals considerable distances.

ionosphere consists of ions (thus the *ion*osphere), which carry a charge. The density of the charge of these ions is heterogeneous in the ionosphere. This is what the RF signals interact with when they enter the ionosphere and this interaction is what causes the signals to be refracted or not. Most of the time an equivalent height is used for calculations involving this form of signal propagation. This height is the height of a layer, which would reflect the signal if it were a plane sheet, rather than refract the signal, which is the actual phenomena involved.

This form of signal propagation is not limited to a single hop. Many hops are possible depending on the conditions of the ionosphere as well as the Earth to where the signal returns. The two-hop case is shown in Figure 4.31.

The direct wave emanating from an antenna in the HF range will travel only so far before its energy gets too small to be useful. The sky-wave signal refracted by the ionosphere returns to the Earth beyond a certain distance as shown in Figure 4.32, creating a *skip zone*. In between these ranges no reception is possible.

4.15.2 Ionospheric Layers

The density of the ions in the ionosphere forms layers, designated by D, E, F1, and F2. However, there occasionally occurs a sporadic E layer, referred to as E_S. These layers are actually bands of ions of similar amounts of charge. Their altitude

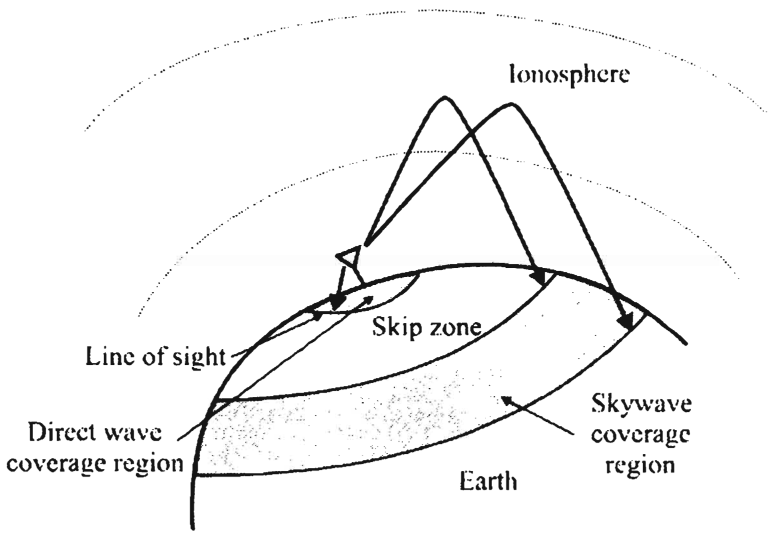


Figure 4.32 Ionospheric propagation creates a skip zone, beyond which the signal returns to the Earth.

depends on circumstances, but the characteristics shown in Figure 4.33 [13] are representative. The altitudes shown in Figure 4.33 correspond to the height of the average electron density for that layer. Signals refracted within these different layers will have different coverage areas on the ground due to their differences in altitude. The electron density of these layers determines whether a signal is refracted or not. Regions with lower electron density will not refract the signals as well as regions with higher densities. Furthermore the refraction properties are frequency-dependent.

4.15.2.1 D Layer

The D layer is located at approximately 60–90 km above the surface of the Earth. Its half-thickness[9] is typically 10 km. This layer is only present during daylight hours and disappears at night. The ionization is highest at noon when the Sun is at its apogee. The D layer is not useful for refracting signals, but it does attenuate them as they traverse through to the higher E and F layers.

4.15.2.2 E Layer

Maximum ionization in the E layer is at about 110 km with a typical half-thickness of 20 km. Like the D layer, the E layer only occurs during daylight hours with its maximum ionization occurring around noon.

[9] Half-thickness is the thickness at which the electron density has dropped to half its maximum.

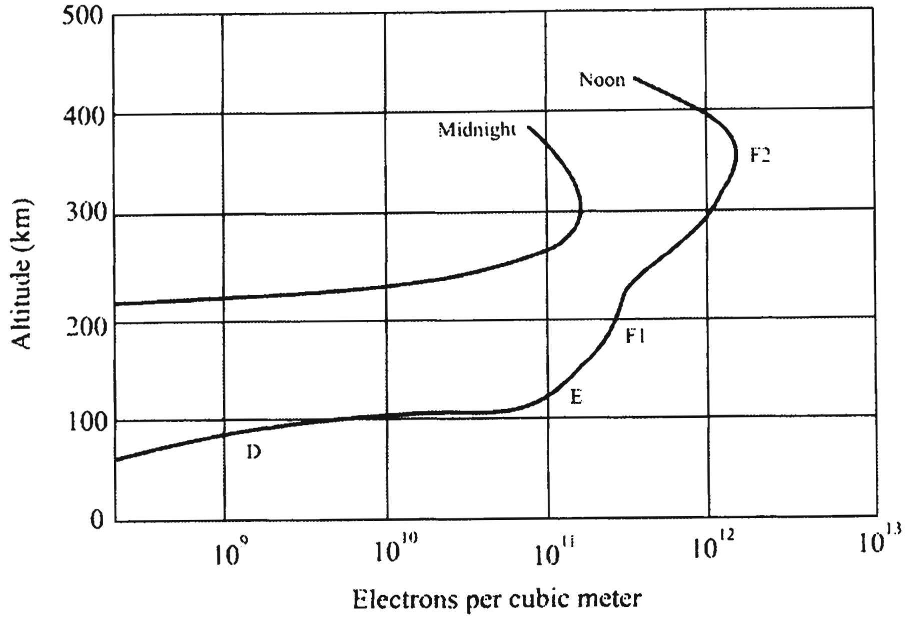


Figure 4.33 Layers in the ionosphere change with time. They determine how HF signals will be refracted. (*Source:* [14]. © Wiley 1986. Reprinted with permission.)

4.15.2.3 E_s Layer

Occasionally there is an ionospheric layer that occurs somewhat higher than the E layer. It is called *sporadic E*, denoted by E_s, and is located at an altitude of 120 km. It typically is very thin, with a half-thickness ranging from a couple hundred meters to about 1 km. The ionization, however, is quite intense.

4.15.2.4 F Layer

The F layer is comprised of two sub-layers, F1 and F2. The F1 layer is located at an altitude of 170–220 km with a half-thickness of typically 50 km. Like the D and E layers, it only occurs during daylight hours. The F2 layer, on the other hand, is present at nighttime as well. It is located at an altitude of 225–450 km and is typically 100–200 km thick.

The two separate F layers only exist during daylight hours. At night, the two layers combine into one—simply the F layer and it is typically located between the altitudes of the F1 and F2 layers. Therefore, the lower frequencies propagate further at night than during the daylight hours, and frequencies that are useable during the day simply pass on through the ionosphere at night.

The D layer discussed above, as well as the E_s layer, cause changing ionospheric conditions, affecting the attenuation characteristics of the ionosphere. This will cause fading at the receiver.

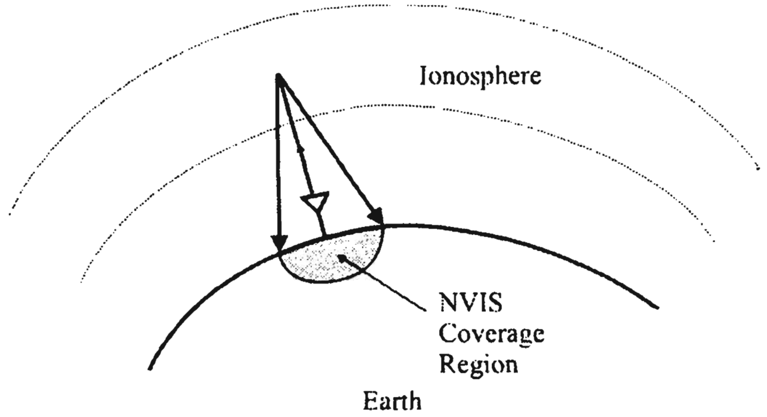


Figure 4.34 NVIS propagation in the HF is for close range. The signal goes straight up and essentially straight back down.

4.15.3 Near Vertical Incidence Skywave

HF signals can be radiated essentially straight up toward the ionosphere by a properly oriented antenna. Under the right conditions these signals will come virtually straight back down, forming a cone, allowing communication within a range of several hundreds of kilometers from the transmit antenna. This mode of communication is referred to as *near vertical incident skywave* (NVIS). The geometry of this mode is illustrated in Figure 4.34, where only half of the ground footprint is shown. To facilitate NVIS communications, the signal must be radiated straight up. This is accomplished by having the antenna arranged horizontally so the boresight of the antenna is pointed straight up. It must also be at the proper height to maximize the signal component radiated straight up.

4.15.4 HF Fading

Fading of EM waves occurs in the HF range. Such fading can be severe, ranging up to 20–30 dB or more. One cause of fading is due to movement of one or more of the nodes. As a node changes position, it changes the point of reception. A node that moves tens of meters changes as much as a wavelength. Since the signal being received typically receives several waves, these waves will add constructively and destructively with movement.

In HF propagation with refraction and reflection via the ionosphere, fading can also occur due to changing and moving ionospheric conditions, whether the communication nodes are moving or not. Fading also occurs because of multiple paths taken through the atmosphere by a signal, adding sometimes destructively and sometimes constructively, to the signal at the receiver.

An ionospheric-refracted wave actually is refracted over a region as opposed to a specific point in space. The index of refraction is not constant over this region, thereby changing the polarization of the wave giving the effect of making the signal fade at the receiver.

4.15.5 Maximum Usable Frequency and Lowest Usable Frequency

The frequencies supported by the ionosphere at any given time form a band. The lowest frequency is termed the *lowest usable frequency* (LUF), and the highest usable frequency is termed the *maximum usable frequency* (MUF). The lower frequency is determined by the ionization caused by the sun during the day. Typical characteristics of the MUF and LUF are shown in Figure 4.35. Depending on frequency and incidence angle, a signal impinging on the ionosphere from the bottom will either be refracted or will pass through, not returning to Earth at all, as illustrated in Figure 4.36.

Ionization of the atmosphere on Earth is caused by solar radiation. The solar radiation is not constant, however, and varies on an approximate 10.7 year cycle. The sunspots are dark areas on the sun, adjacent to which are typically flares of intense radiation. These flares affect the Earth's ionosphere by increasing the amount of ionization. The result is higher critical frequencies, especially in the highest layer, the F2 layer. This, in turn, facilitates longer range communications during peaks of solar activity. In addition the sun rotates on an axis, and the radiation activity around the sun is not constant. A region of high radiation facing the Earth will cause additional ionization, which rotates away. This period of rotation is approximately 27 days.

4.15.6 Automatic Link Establishment

The aforementioned anomalies of communicating in the HF range illustrate some of the difficulties of using the high frequency range for communication purposes. This difficulty is determining the proper frequency to use, which, as just mentioned, depends on the state of the ionosphere. In the past, it has required skilled technicians to make this determination, thus relegating usage of the HF frequency range to technical experts. With *automatic link establishment* (ALE) techniques, the selection of the proper frequency is performed by the radio equipment itself, precluding the requirement for the operator to have extensive technical experience at such propagation and equipment operation.

The equipment does this by probing the ionosphere with signals of varying frequency and determining the frequency that works the best. It determines this by measuring the quality of the signals received at each frequency. Each end of the link performs the signal quality measurement.

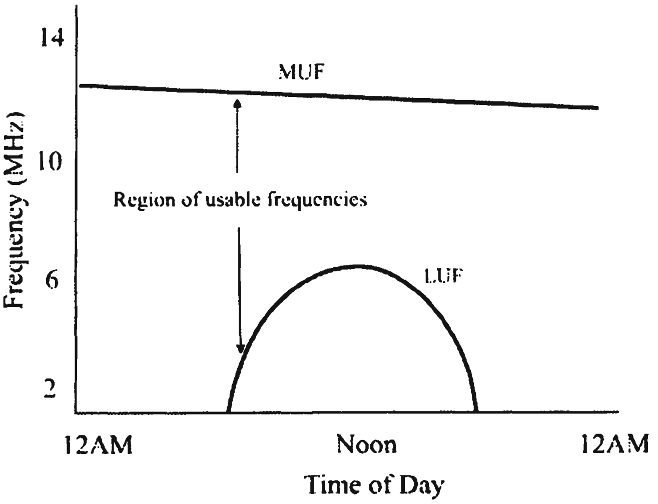


Figure 4.35 Maximum frequency and minimum frequency of ionospheric propagation.

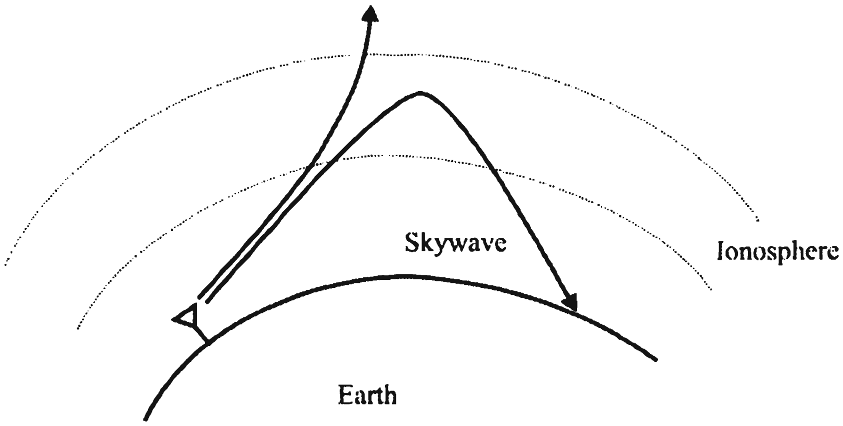


Figure 4.36 An HF signal at the wrong frequency and/or wrong attack angle will traverse through the ionosphere without coming back to Earth. This is the primary cause of the skip zone.

The overall technique for ALE is as follows:

- Each station in the network is assigned an address as a call sign consisting of 1 to 15 characters.
- Each station in the network monitors a pre-determined number of frequencies, say 2–100 at a specified scan rate, say 2–10 frequencies per second.
- Each station in the ALE network broadcasts its call sign on each frequency at a specified interval.
- If a station hears its call sign, while scanning, broadcast by another station, it can respond. No matter if it responds or not, it will record the *link quality analysis* (LQA) value.
- If a station desires to contact another station it will select the frequency to the station with the largest LQA value for that station or by calling on each frequency in the scan list.
- If a station needs to contact a single station, a group of stations, the entire network, or a special subset of the network either by call sign or portion of call sign, (e.g., all members with "AA" in the call sign), it will broadcast a message specifying the callsign of the desired station(s), or special call signs indicating all, or any station, on one of the assigned network frequencies.
- If a station receives a call or it if places a call and the appropriate station(s) responds the station will notify the operator of the link.

Thus ALE allows any desired number of stations to communicate as required without an operator being involved with network operations.

4.16 Concluding Remarks

Propagation of radio waves is facilitated by different mechanisms depending largely on the frequency of the signal. Lower frequencies tend to propagate farther than higher frequencies all else being equal. The propagation characteristics of signals that are normally of interest to military communication were presented in this chapter. These characteristics are of similar importance to designers of communication systems as well as communication EW systems— the propagation effects are on the signals themselves, irrespective of what the signals are being used for.

Some characteristics, however, perturb the operation of communication systems vis-à-vis communication EW systems. An example of this is meteor burst communications. One of the more important measurable parameters of communication signals of interest to EW systems is the location of the target. This is unobtainable in meteor burst systems since the transmitter sends the signal toward the meteors, not the receiver.

The ionosphere will reflect (refract) HF waves so very long distance communications are possible. Although there are exceptions, this form of communication is not possible much above the HF range, however. For the higher frequency range, propagation is normally limited to radio line of sight (4/3 Earth model), which significantly limits the distance over which two nodes can communicate close to the Earth. For communication EW systems, this normally means that to intercept such signals, it is normally more efficient to elevate the receiving antenna. Such elevation typically means putting the receiving antenna on some form of aircraft.

References

[1] Hall, M. P .M., *Effects of the Troposphere on Radio Communication*, Stevenege, UK: Peter Peregrinus, Ltd., 1979.

[2] Shibuya, S., *A Basic Atlas of Radio Wave Propagation*, New York: Wiley, 1987.

[3] Parsons, D., *The Mobile Radio Propagation Channel*, New York: Wiley, 1994.

[4] Stuber, G. L., *Principles of Mobile Communications*, Boston: Kluwer Academic Publishers, 1996.

[5] Gagliardi, R. M., *Introduction to Communications Engineering*, New York: Wiley, 1988.

[6] *Reference Data for Radio Engineers*, Ch. 28, Indianapolis, IN: Howard W. Sams & Co., Inc., 1975.

[7] Davies, K., *Ionoshperic Radio*, London: Peter Peregrinus Ltd., 1989.

[8] Hall, M. P. M., *Effects of the Troposphere on Radio Communication*, Stevenege, UK: Peter Peregrinus, Ltd, 1979, p. 1.

[9] Rappaport, T. S., *Wireless Communications: Principles and Practice*, Upper Saddle River, NJ: Prentice Hall, 1996.

[10] Schanker, J. Z., *Meteor Burst Communications*, Norwood, MA: Artech House, 1990.

[11] Parsons, D., *The Mobile Radio Propagation Channel*, New York: Wiley, 1992, pp. 134–136.

[12] Wojnar, A. H., "Unknown Bounds on Performance in Nakagami Channels," *IEEE Transactions on Communications*, Vol. COM-34, No. 1, January 1986, pp. 22–24.

[13] Braun, G, *Planning and Engineering of Shortwave Links*, Siemens Aktiengesellschaft, New York: Wiley, 1986, p. 21.

[14] Davis, K., "Ionospheric Radio Propagation," NBS Monograph 80, U.S. Government Printing office, 1965.

Chapter 5

Radio Frequency Noise and Interference

5.1 Introduction

Noise is the nemesis of all electronic systems. At any temperature above absolute zero, agitated electrons move at random in electronic components which produce stochastic currents. Additional effects in active devices such as semiconductors and vacuum tubes produce random noise currents as well. Lightning bolts in thunderstorms produce large bursts of random signals. These are just some of the noise sources we will discuss in this chapter.

Noise in the higher frequency ranges is caused by several mechanisms. These sources of noise typically are not limited to the high frequency ranges, but extend down to almost DC. Our interest here is to discuss the cause and effects of noise in the RF range.

The taxonomy of noise sources can be dissected into two broad areas: those noise sources that are external to the electronic system in question and those that are internal to that system. The external noise sources are, for the most part, uncontrollable by the designer of these systems, but their effects must be dealt with as best as possible. On the other hand, with careful consideration, the effects of internal noise sources can be minimized to optimize the performance of EW systems.

There is another type of phenomena that is sometimes grouped with noise. This is the spurious response of electronic systems due to nonlinearities. While having detrimental effects on the response of such systems, the mechanisms for generating the spurious responses are fundamentally different. Spurious signals are not discussed further here.

Noise from any source can be extremely detrimental to the operation of EW systems. The receivers are designed to be very sensitive and they can pick up even minor noise. Such noise defines the noise floor, below which intercept is normally very difficult. Raising the noise floor tends to reduce the intercept range. Noise also causes excess errors to enter into the geolocation calculations because the angle or time measurements are not as precise as otherwise. Lastly, EW systems must be tasked via C2 and report their results via RF communications. Excess noise can limit the deployment options due to limiting the lengths of these C2 links.

This chapter discusses the predominant noise sources of concern to communication systems, and in particular, communication EW systems. It is organized as follows. Section 5.1 presents a discussion of the dominant sources of external radio noise. The material in this section is derived largely from [1]. Section 5.2 puts forth the predominant sources of noise generated internally in electronic systems. The internal sources are the cause of, arguably, the most important internal noise source in modern communication EW systems: phase noise in oscillators.

5.2 Sources of External Radio Noise

Radio noise external to a radio receiving system originates from the following sources [1]:

- Radiation from lightning discharges (atmospheric noise due to lightning);
- Unintended radiation from electrical machinery, electrical and electronic equipments, power transmission lines, or from internal combustion engine ignition (man-made noise);
- Emissions from atmospheric gases and hydrometeors;
- The ground or other obstructions within the antenna beam; and
- Radiation from celestial radio sources.

Noise or signals due to unwanted cochannel transmissions or due to spurious emissions from individual transmitting or receiving systems are also noise sources, but they are not included in the discussion in this chapter. These topics are difficult to treat in general as they are always situation-dependent.

There are several noise generators external to an EW system. These sources vary depending on the frequency under consideration. Man-made noise comes from several sources that are generated by machinery or other man-made devices. Examples of these sources are automobile ignitions, welding machines, and microwave ovens. Clearly the amount of this noise will depend on the number of such interference sources present. More noise will be generated in manufacturing settings than on farms, for example. More man-made noise will be found in cities than in the country.

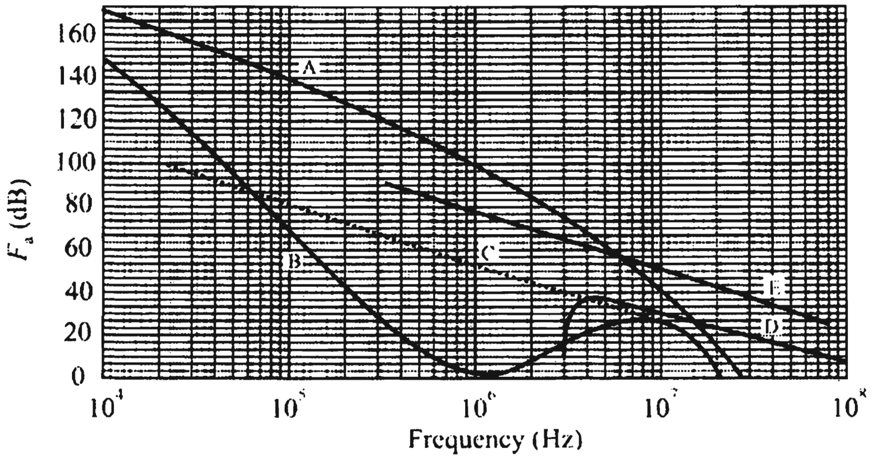


Figure 5.1 F_a for low VHF; low RF noise; A: atmospheric noise, value exceeded 0.5% of the time; B: atmospheric noise, value exceeded 99.5% of the time; C: man-made noise, quiet receiving site; D: galactic noise; and E: median business area man-made noise; minimum noise level expected. (*Source:* [2]. © Artech House 2004. Reprinted with permission.)

The atmosphere surrounding the Earth contains a certain amount of heat energy at any given time. This heat energy warms up the electrons in the air, which in turn radiate a certain amount of thermal noise. This noise is picked up by the antennas in a communication system and is manifest as thermal noise at the system inputs. Compared to other noise sources, this one contributes in a relatively minor way.

Noise is also generated by energy sources (stars) in the universe, including the Earth's own Sun [1]. If an antenna is pointed toward a star, for example, then that star will cause significant broadband energy to be introduced into the system. These noise sources tend to be wideband, containing noise energy across a significant portion of the spectrum. This form of noise is sometimes referred to as *galactic noise*.

Another principal cause of *atmospheric noise* is lightning strikes from thunderstorms. This source of noise is prevalent in the HF frequency range and therefore can propagate for significant distances. Thousands of thunderstorms occur each year with many more thousands of lightning bolts. Thus, this type of noise is a problem almost everywhere.

The amount of noise generated by these sources is typified by the data shown in Figures 5.1–5.5, where the variation with frequency is obvious. The total amount of noise present is given by a multiplicative factor on *kTB* as above. Thus,

$$\begin{aligned} N_{\text{total}} &= N \times N_{\text{external}} \\ &= kTBFN_{\text{external}} \end{aligned}$$

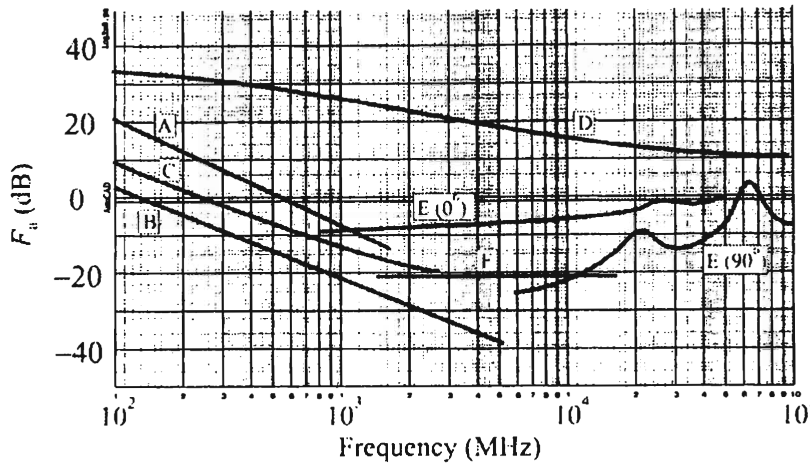


Figure 5.2 High RF noise; A: estimated median business areas man-made noise; B: galactic noise; C: galactic noise toward galactic center with infinitely narrow beamwidth; D: quiet sun (1/2 degree beamwidth directed at sun); E: sky noise due to oxygen and water vapor (very narrow beam antenna), upper curve, 0° elevation angle, lower curve, 90° elevation angle; F: black body (cosmic background), 2.7 K, minimum noise level expected. (*Source*: [2]. © Artech House 2004. Reprinted with permission.)

$$\frac{N_{\text{total}}}{kTB} = FN_{\text{external}}$$

$$\begin{aligned} N_{\text{total, dBkTB}} &= 10\log\frac{N_{\text{total}}}{kTB} = 10\log F + 10\log N_{\text{external}} \\ &= F_{\text{dB}} + N_{\text{external, dBkTB}} \qquad (5.1) \end{aligned}$$

So the total noise in decibels above *kTB* is given by the noise figure in decibels plus the external noise in decibels relative to *kTB*. That is why the ordinate in these figures is in the units indicated.

The external noise factor for the frequency range 10 kHz to 100 MHz for various categories of noise is shown in Figure 5.1 [1]. For atmospheric noise, the minimum values of the hourly medians expected for atmospheric noise are those ordinate values exceeded 99.5% of the time and the maximum values are those exceeded 0.5% of the time. All times of day, seasons, and the entire Earth's surface have been taken into account for these atmospheric noise curves; thus the wide variations. Similarly, Figure 5.2 [2] covers the 100 MHz to 100 GHz frequency range.

The majority of the results shown in the two figures are for omni-directional antennas (except as noted on the figures). The results for directional antennas, however, can vary from these curves. At HF, for example, with atmospheric noise due to lightning strikes and with narrow beam antennas, there can be as much as

Table 5.1 Man-Made Site Definitions

Category	Definition
City center	Center of large city in proximity to light industry, offices and public transport systems (road/train/underground)
Factory estate	Concentration of industrial units and factories undertaking light to medium industrial activities
Business center	Modern business center containing a concentration of office automation equipment (PCs, fax machines, photocopiers, telephones, etc.)
Suburban	Mainly residential area on the outskirts of a town/city
Rural	Countryside location, but with evidence of human activity (small groups of houses and shops, minor roads)
Quiet rural	Countryside location, but with little or no evidence of human habitation

Source: [1].

10 dB variation (5 dB above to 5 dB below the average F_a value shown) depending on antenna pointing direction, frequency, and geographical location.

The average value (averaged over the entire sky) of galactic noise is given by the curve labeled galactic noise in Figures 5.1 and 5.2. Measurements indicate a ±2 dB variation about this curve. The minimum galactic noise (narrow beam antenna toward galactic pole) is 3 dB below the solid galactic noise curve shown in Figure 5.2.

5.2.1 Man-Made Noise

Man-made noise is generally divided into two types of sources [1]. The first is WGN similar to that discussed above. The second is *impulsive noise* (IN) that is characteristic of certain types of sources such as automobile ignitions. The noise levels vary according to sighting criteria—noise in a city setting is higher than the noise on a farm, for example. Table 5.1 gives the definitions of the sites for the noise measurements presented here [1, 3–5].

5.2.1.1 WGN

The external noise figure, F_a, as defined earlier describes the level of Gaussian noise in the environment due to man-made sources. Typical results are shown in Figure 5.3. These curves follow

$$F_a = c - d \log f \tag{5.2}$$

where F_a is in dB and f is in MHz. Values for c and d are given in Table 5.2.

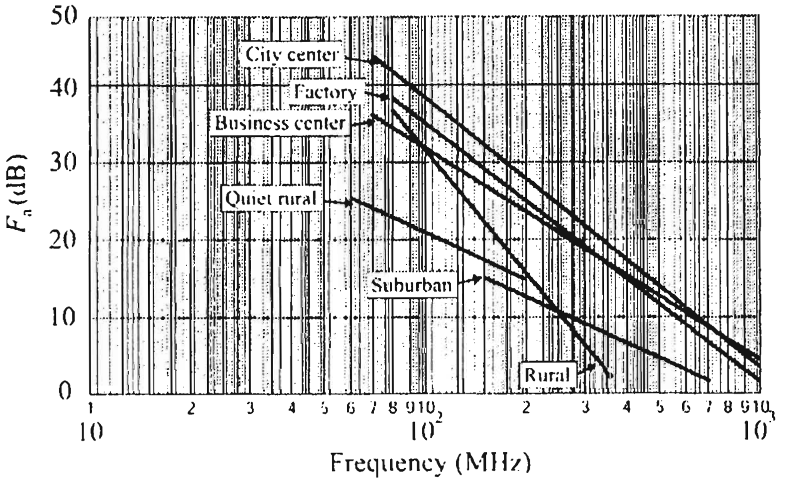


Figure 5.3 External Gaussian noise due to man-made sources. (*Source:* [1].)

Table 5.2 Values of *c* and *d* for WGN in (5.2)

Noise category	***c***	***d***
City center	111.6	36.1
Factory estate	104.5	34.4
Business center	89.5	28.4
Suburban	60.8	20.8
Rural	141.0	54.5
Quiet rural	62.3	20.6
Galactic noise	52.0	25.0

Source: [1].

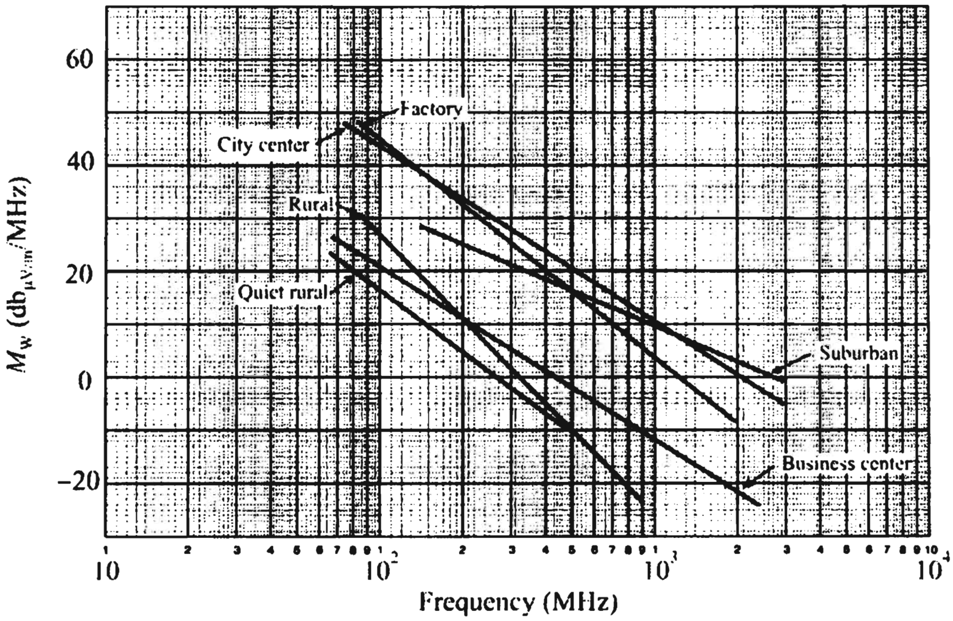


Figure 5.4 Mean impulsive noise voltage. (*Source:* [1].)

5.2.1.2 Impulsive Noise

The man-made noise due to impulsive sources is given in terms of a voltage level (actually in dB relative to 1 μV/MHz) rather than a power level. This is due to the way the data was collected. Figure 5.4 shows representative levels of the mean IN voltage, M_w in dB(μV/MHz), plotted versus frequency which follows a similar relationship to that of WGN:

$$M_W = g - h \log f \tag{5.3}$$

where f is in MHz. The coefficients for (5.3) are shown in Table 5.3.

Table 5.3 Values of g and h for Impulsive Noise in (5.3)

Environmental category	g	h
City center	79.3	32.9
Factory estate	96.8	40.8
Business center	56.6	32.6
Suburban	45.9	21.9
Rural	104.1	55.3
Quiet rural	65.7	39.3

Source: [1].

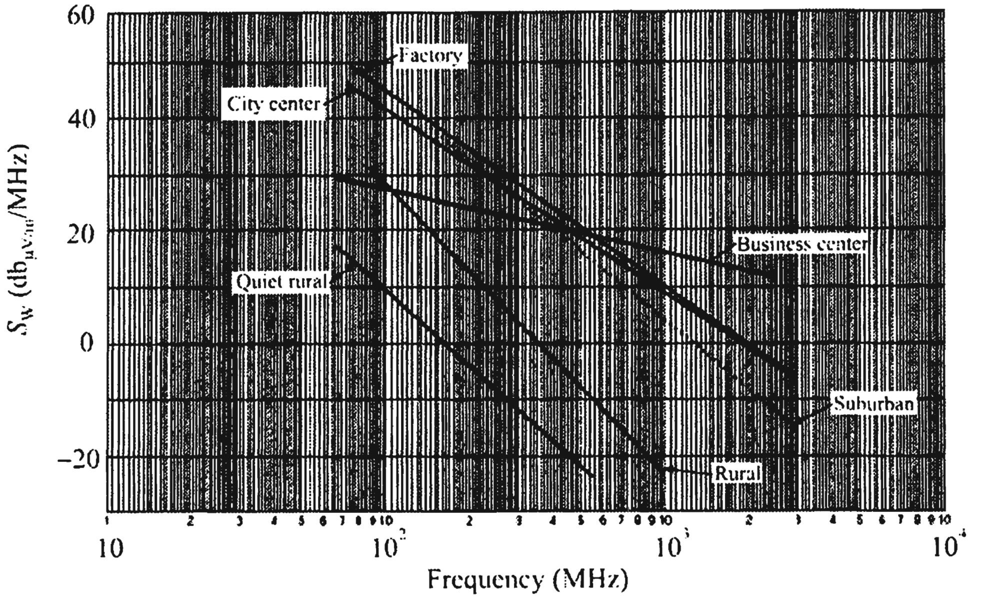


Figure 5.5 Standard deviation of the mean IN voltage. (*Source:* [1].)

Standard Deviation of Impulsive Noise Voltage

Typical characteristics of the standard deviation of IN, S_W, versus frequency are illustrated in Figure 5.5. In most cases there is a good linear relationship such that S_W in dB(μV/MHz) falls with log frequency yielding the relationship

$$S_W = m - n \log f \tag{5.4}$$

when f is in MHz. The coefficients for (5.4) are given in Table 5.4. Expression (5.4) is sensitive to the number of impulses detected.

5.2.1.3 Noise Due to Extra-Terrestrial Sources

Objects in the sky are sources of RF noise. The noise levels from such sources are usually given in terms of a brightness temperature, t_b. The antenna temperature, t_a,

Table 5.4 Values of m and n for the Standard Deviation of the Impulsive Noise in (5.4)

Environmental category	m	n
City center	106.5	32.4
Factory estate	116.3	35.6
Business center	50.8	11.5
Suburban	126.9	41.0
Rural	130.4	51.2
Quiet rural	100.9	45.6

Source: [1].

is the convolution of the antenna pattern and the brightness temperature of the sky and ground. For antennas whose patterns encompass a single source, the antenna temperature and brightness temperature are the same (curves C, D, and E of Figure 5.2, for example).

As a general rule, for communications below 2 GHz, we need to be concerned with the Sun and our Milky Way galaxy, which appears as a broad belt of strong emission. For $f < 100$ MHz, the median noise figure for galactic noise, neglecting ionospheric shielding, is given by

$$F_{\text{am}} = 52 - 23\log f \tag{5.5}$$

where f is the frequency in MHz. Above 2 GHz, we need consider only the Sun and a few very strong sources as indicated in Table 5.5 since the cosmic background contributes only 2.7 K and the Milky Way appears as a narrow zone of somewhat enhanced intensity. The brightness temperature range for the common extraterrestrial noise sources in the frequency range 0.1 to 100 GHz is illustrated in Figure 5.6.

5.2.1.4 Unintentional and Intentional EMI

Just as *electromagnetic interference* (EMI) can be a source of noise internal to a system, it can be a source of unwanted noise external to a system. Systems within close proximity of an EW system can radiate signals that interfere with the proper operation of that system. This can be a particularly significant problem for EW systems because such systems are typically broadband relative to, say, communication systems. Broadband systems are more susceptible to a broader range of EMI than many other systems. In addition, EW systems are normally

Table 5.5 Characteristics of Galactic Noise Sources

Source	Temperature, K	Beamwidth, degrees
Sun	6,000	0.5
Moon	200	0.5
Stars (at 300 MHz)		
Cassiopeia	3,700	$<10^{-3}$
Cygnus	2,650	$<10^{-3}$
Taurius	710	$<10^{-3}$
Centaurus	460	$<10^{-3}$
Planets (10 MHz–10 GHz)		
Mercury	613	2×10^{-3}
Venus	235	6×10^{-3}
Mars	217	4.3×10^{-3}
Jupiter	138	1.3×10^{-3}
Saturn	123	5.7×10^{-3}
Earth (from moon)	300	2

Source: [6].

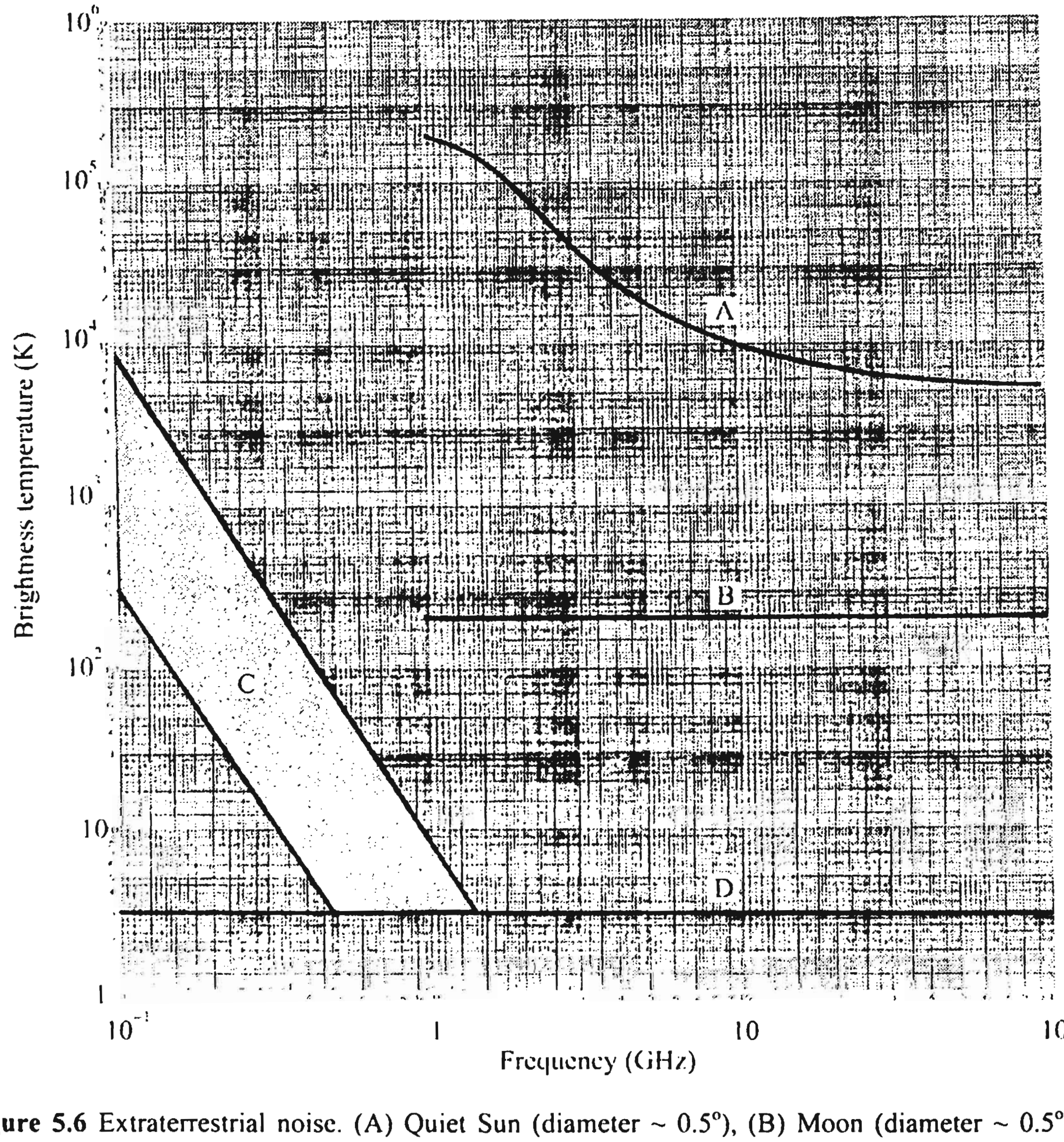


Figure 5.6 Extraterrestrial noise. (A) Quiet Sun (diameter ~ 0.5°), (B) Moon (diameter ~ 0.5°); (C) Range of galactic noise; and (D) Cosmic background. (*Source:* [1].)

made as sensitive as possible so that distant signals can be received. This makes them even more susceptible to EMI.

Intentional jamming of an EW system can be considered as noise and can be treated the same. This type of jamming is sometimes used to screen friendly communications from intercept by an adversary's communication EW systems. This is a form of EP, another category of IW.

5.2.2 External Noise Summary

There are several sources of noise external to an EW system. Included in this are atmospheric noise originating predominantly by lightning discharges, man-made noise emanating from any device that is man-made, and celestially originated emanations such as from stars. The level of these noise sources varies with location, time-of-year, and many other factors. There is very little that can be done to avoid this noise, and EW systems must accommodate them.

5.3 Internal Noise in EW Systems

5.3.1 Introduction

Noise is always present in communication systems and the sensors that target them. Every electronic device generates internal noise due to the excitation of electrons when the temperature is raised above absolute zero—this is the thermal noise and it is not just within the system itself. Thermal noise exists in the atmosphere as well as will be described. This added noise decreases the quality of the signal and makes subsequent signal processing, such as demodulation, more difficult.

There are several types of internal noise sources that are included in electronic systems [7–13]:

1. Thermal noise (Johnson or Nyquist noise),
2. Shot noise,
3. Low frequency noise (flicker noise),
4. Generation-recombination noise in semiconductors,
5. Oscillator phase noise,
6. A/D conversion noise, and
7. Noise radiated from one component to another—(EMI).

Each of these sources of noise is discussed in this section.

5.3.2 Broadband Noise Sources

Thermal noise and shot noise are both "white" noise sources, i.e., power per unit bandwidth (which is also called the spectral density) is constant:

$$\frac{dP_{\text{noise}}}{df} = N_0 \tag{5.6}$$

5.3.2.1 Thermal Noise

In Planck's theory of black body radiation, the radiated energy is given by

$$\bar{E} = \frac{h\upsilon}{e^{h\upsilon/kT} - 1} \tag{5.7}$$

and the spectral density of the radiated power is given by

$$\frac{dP}{d\upsilon} = \bar{E} \tag{5.8}$$

so that

$$\frac{dP}{d\upsilon} = \frac{h\upsilon}{e^{h\upsilon/kT} - 1} \tag{5.9}$$

P is the power that can be extracted in equilibrium.

At room temperature and low frequencies (less than 100 GHz), $h\nu << kT$, so the exponential can be approximated by the first two terms of its Taylor series expansion with negligible error

$$e^{h\upsilon/kT} \approx 1 + \frac{h\upsilon}{kT} \tag{5.10}$$

so that

$$\frac{dP}{d\upsilon} \approx \frac{h\upsilon}{1 + \frac{h\upsilon}{kT} - 1} = kT \tag{5.11}$$

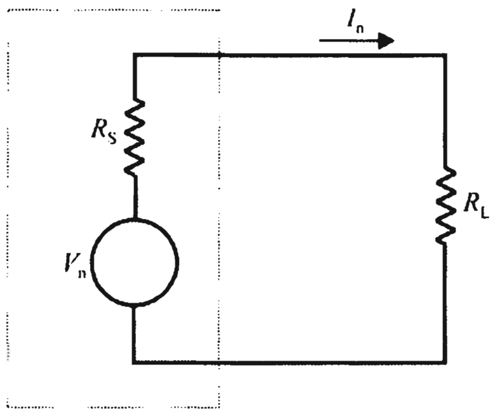


Figure 5.7 Resistive noise source. The dotted box encloses the equivalent circuit of the resistive noise source.

so at low frequencies the spectral density is independent of frequency. Furthermore, for a total bandwidth B the noise power that can be transferred to an external device is given by

$$P_n = \int_0^B dP = \int_0^B kTd\upsilon = kTB \tag{5.12}$$

To apply this result to the noise of a resistance (which could be the resistance of a resistor, capacitor, inductor, or any other circuit element), consider a resistor R whose thermal noise gives rise to a noise voltage V_{n} as illustrated in Figure 5.7. This resistance is connected to an external resistor causing a current I_{n} to flow through the load resistor. The power dissipated in the load resistor R_{L} is given by

$$\frac{V_{\mathrm{n}R_{\mathrm{L}}}^2}{R_{\mathrm{L}}} = \frac{V_{\mathrm{n}}^2 R_{\mathrm{L}}}{(R+R_{\mathrm{L}})^2} \tag{5.13}$$

As is well known, maximum power transfer occurs when the load resistance equals the source resistance $R_{\mathrm{L}} = R$,

$$V_{\mathrm{nR_L}}^2 \Big|_{\text{maximum power transfer}} = \frac{V_{\mathrm{n}}^2}{4} \tag{5.14}$$

From (5.12), the maximum power that can be transferred to R_{L} is kTB so that

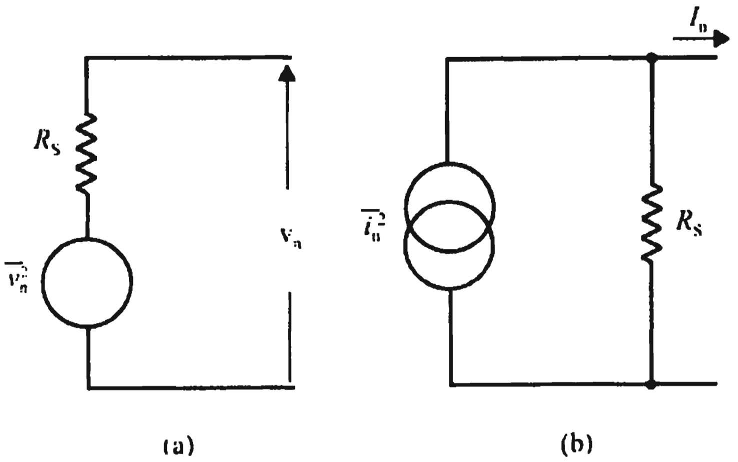


Figure 5.8 Equivalent circuits of a thermal noise source. (a) Thevenin and (b) Norton.

$$\frac{V'^2_{nR_L}}{R} = \frac{V'^2_n}{4R} = kTB \tag{5.15}$$

and

$$P_n = \frac{V'^2_n}{R} = 4kTB \tag{5.16}$$

and the spectral density of the noise power in the resistor is given by

$$\frac{dP_n}{dB} = 4kT \tag{5.17}$$

In this approximation, v_n is independent of frequency. For this reason, the thermal noise signal is called "white noise." The noise voltage is modeled as a random variable with a zero mean Gaussian distribution with variance $\overline{v_n^2}$. Given multiple resistive components, and therefore noise sources, the distributions are normally considered to be independent. This means that if you combine multiple noise sources, the variance of the sum is equal to the sum of the variances (we add the noise powers, not the voltages).

We can replace any noisy (warm) resistor with a Thevenin or Norton equivalent of a noise source and an ideal, noiseless resistor as seen in Figure 5.8. If we connect this equivalent circuit to a bandpass filter with bandwidth B Hz and then to a second ideal resistor R (where the resistance of the load is chosen for maximum power transfer), the noise power delivered to the load is

$$P_{\mathrm{n}} = \left(\frac{\overline{v}_{\mathrm{n}}}{2R} \right)^2 R = \frac{\overline{v}_{\mathrm{n}}^2}{4R} \tag{5.18}$$

Note that we do not have another two in the denominator since the voltage is already an RMS quantity. Now

$$\overline{v_{\mathrm{n}}^2} = 4k_{\mathrm{B}}TBR \tag{5.19}$$

so

$$P_{\mathrm{n}} = \frac{4k_{\mathrm{B}}TBR}{4R} = k_{\mathrm{B}}TB \tag{5.20}$$

5.3.2.2 Shot Noise

If an excess electron is injected into a device, it forms a current pulse of duration τ consisting of a single charge, but nevertheless a current [14]. In a vacuum tube τ could be the transit time from cathode to anode. In a semiconductor diode τ could be the recombination time (the time it takes for the electron to recombine with an atom). If these times are short with respect to the periods of interest ($\tau << 1/f$), the current pulse can be modeled by an impulse of unit strength. The Fourier transform of an impulse is a "white" spectrum, i.e., the amplitude distribution in frequency is uniform. It is called white because for all intents and purposes all frequencies are present, just as white light consists of all colors (wavelengths/frequencies). If N electrons are emitted at the same average rate, but at different times, they will have the same spectral distribution, but the coefficients will differ in phase. For example, for two currents i_{a} and i_{b} with a relative phase ϕ, the total RMS current is

$$< i^2 > = (i_{\mathrm{a}} + i_{\mathrm{b}}e^{j\varphi})(i_{\mathrm{a}} + i_{\mathrm{b}}e^{-j\varphi}) = i_{\mathrm{a}}^2 + i_{\mathrm{b}}^2 + 2i_{\mathrm{a}}i_{\mathrm{b}}\cos\phi \tag{5.21}$$

For random phase the third term averages to zero so

$$< i^2 > = i_{\mathrm{p}}^2 + i_{\mathrm{q}}^2 \tag{5.22}$$

Extending this to N electrons randomly emitted per unit time, the individual spectral components simply add in quadrature for a total current of

$$i_\mathrm{n}^2 = 2Nq_\mathrm{e}^2 \tag{5.23}$$

where q_e is the charge of an individual electron, $q_\mathrm{e} = 1.6022 \times 10^{-19}$ C. The average, or DC, current is

$$I = Nq_\mathrm{e} \tag{5.24}$$

so $N = I / q_\mathrm{e}$ and the spectral noise density is

$$i_\mathrm{n}^2 = \frac{dI_\mathrm{n}^2}{df} = 2q_\mathrm{e}I \tag{5.25}$$

where I is the DC current through the device.

Shot noise does not occur in "ohmic" conductors because the number of available charges is essentially unlimited and the fields caused by local fluctuations in the charge density draw in additional carriers to equalize the total number.

5.3.2.3 Broadband Noise Power

The expression for the thermal noise power in a resistance and the shot noise in a semiconductor are independent of frequency. As noted above, the psds for such noise sources are constant levels. Therefore the amount of noise *power* present depends on the bandwidth under consideration. The total noise increases with bandwidth as illustrated in Figures 5.9–5.11—the total noise is the integral over the noisy region. When the noise is broadband as illustrated in Figure 5.9, considerable amounts of unwanted noise are present with the signal. A significantly reduced noise power is illustrated in Figure 5.10, where the noise is baseband. Nevertheless, there is a considerable amount of noise power present. The ideal situation is illustrated in Figure 5.11 where the noise bandwidth is minimized—it's just large enough to pass the signal unperturbed. The SNR is decreased as the excess bandwidth is reduced until the signal becomes substantially affected.

5.3.3 Low-Frequency Noise

In many physical phenomena at low frequencies there is a characteristic noise with a spectrum proportional to $1/f^\alpha$ with α on the order of one as illustrated in Figure 5.12 [15–19]. Interestingly, when α = 0, this noise characteristic corresponds to thermal noise. When α = 1, the noise is called *1/f noise,* or *flicker noise* and *pink*

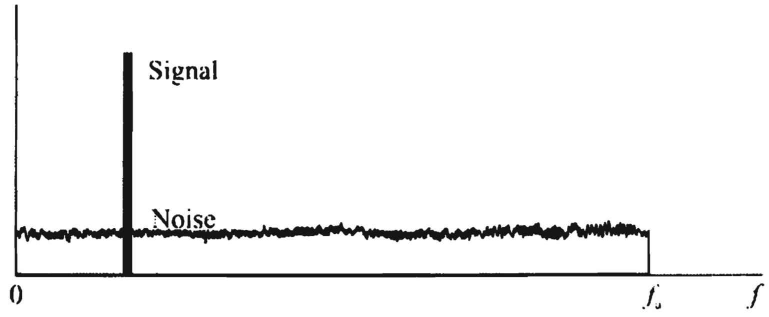


Figure 5.9 Spectrum of broadband noise. Much more noise is present than necessary to process the information in the signal.

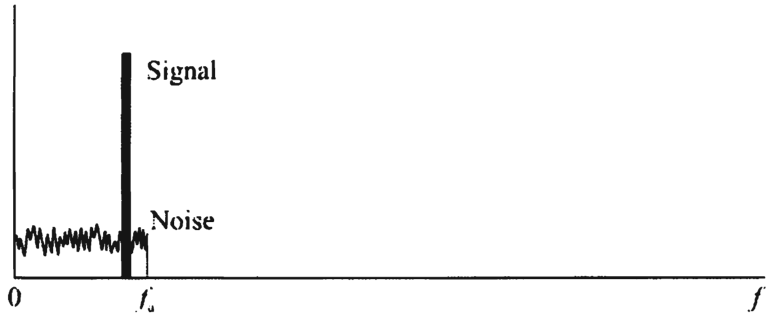


Figure 5.10 Spectrum of low-pass noise. There is much less noise than in Figure 5.9, but still more than necessary.

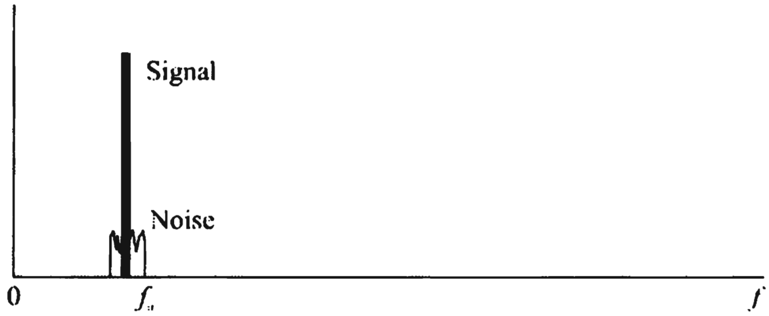


Figure 5.11 Spectrum of bandpass noise. In this case the bandpass filter limits the noise to only that necessary to pass the signal unfettered.

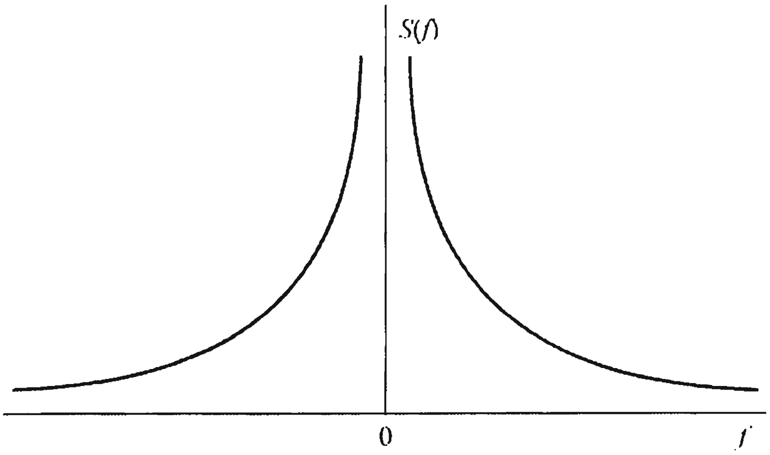


Figure 5.12 Spectrum of low-frequency noise behaves as $1/f^{\alpha}$ where $\alpha \sim 1$.

noise. When $\alpha = 2$, the characteristics of Brownian noise are observed [20]. This noise is typically measured over a considerable frequency range, but 1 Hz to 10 kHz is typical. Note that the spectrum cannot be exactly f^{-1} from $f = 0$ to $f = \infty$, since neither the integral of the power density nor the Fourier transform would have finite values. At some higher frequency f_h the slope must be steeper than −1. However, this f_h has never been observed because the $1/f$ noise disappears into the white thermal noise that is always present below that frequency. There also should be a lower limit, below where the noise spectrum flattens. However, measurements as low as 10^{-6} Hz showed that even there the spectrum still is f^{-1} [21].

5.3.4 Generation-Recombination Noise

In semiconductors, there is a static charge carrier generation/recombination process that generates noise. This noise is also called *burst-noise* or *popcorn-noise* [14]. Generation-recombination noise is constant at low frequencies, and past a certain frequency point, it falls proportionally to $1/f^2$ as illustrated in Figure 5.13. The generation-recombination (g-r) process is a natural part of all semiconductor behavior. In the semiconductor, carriers are freed from association with a particular atom by a generation process, which is necessary for either of the conduction mechanisms, drift and diffusion, to occur. This leaves ionized donors or acceptors which will occasionally trap a passing carrier. Because of the thermal energy in the crystal lattice, the trapped carrier will be freed again after only a short time. This process is a series of independent discrete events. Each event causes fluctuation in the number of free carriers leading to a fluctuation in the material resistance. If DC is passed through the material, a fluctuating voltage related to the fluctuating resistance will appear across the device. This generation-

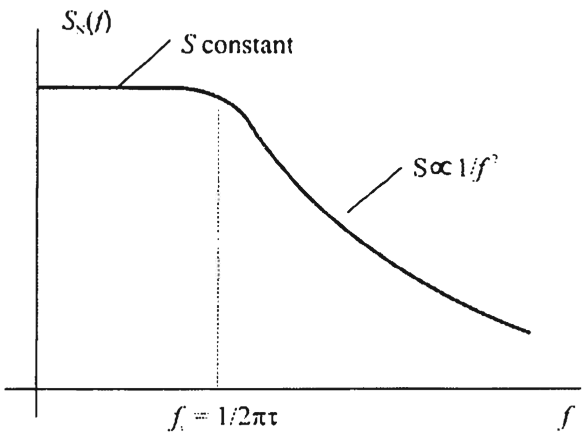


Figure 5.13 A single time constant spectrum known as a *Lorentzian* has one sharp corner defined by the time constant.

recombination noise: (1) will not appear if there is no DC; (2) is a function of current; however, it is *not* produced by the current, and (3) is a low-frequency noise. This mechanism generates a Cauchy spectral distribution, which some literature calls a Lorentzian distribution, which is illustrated in Figure 5.8. Different burst-noise mechanisms within the same device can exhibit different corner frequencies. When superimposed on the flicker noise, burst noise can cause bumps in the flicker noise's otherwise-straight spectral slope.

The spectral density of the number of carriers, N, in a semiconductor is given approximately by

$$S_N(f) \approx 4\frac{N_0 P_0}{N_0 + P_0}\frac{\tau}{1+\omega^2\tau^2} \tag{5.26}$$

where

$\omega = 2\pi f$
N_0 = equilibrium value of N, a constant
$P_0 = N_0 - N_d$, a constant
N_d = donor concentration in the semiconductor, a constant
τ = time constant of the generation-recombination process, a constant

Equation (5.26) gives a relation for the spectral density of the carrier density fluctuation and exhibits the characteristics of a Lorentzian distribution shown in Figure 5.13. The frequency response is constant at low frequency with a corner at a frequency $f_c = 1/2\pi\tau$. Above this, the frequency slope is proportional to $1/f^2$.

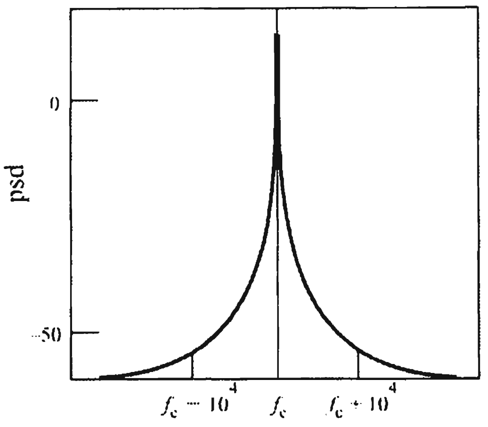


Figure 5.14 Spectrum of phase noise.

5.3.5 Phase Noise

Phase noise of oscillator circuits is among the key parameters of modern communication systems [22–26]. It is not a noise source per se, as thermal noise and shot noise are. It is caused by changes in the operating frequency of an oscillator due to the other types of noise sources. Phase noise limits the modulation quality of the information signal and the signal spill-over into to adjacent channels. In the receiver it can reduce the selectivity and the demodulation quality. With today's trend to even more complex modulation schemes such as higher order *phase shift key* (PSK), low-phase noise becomes even more critical. Therefore, minimizing noise with the design of high-frequency communication circuits is a must for the cost-effective exploitation of limited bandwidth, and to obtain low bit error rates.

Phase noise has considerable impact on EW system performance. Suffice it for now to say that every oscillator exhibits a psd characteristic similar to that shown in Figure 5.14, where there are frequency components in the oscillator signal that are not always exactly the designed frequency.

A key element is the transistor. Active devices are inherently nonlinear at some point in their operation characteristics, and nonlinearities process noise differently from the way linear circuits do. While for linear circuits the modeling of the noise at the operating frequency f_0 is frequently sufficient, nonlinear circuits convert the low-frequency noise up to the operating frequency range. For oscillators, as an example, the $1/f$ low-frequency noise at f_m will be mixed upwards to contribute to the phase noise of the total circuit at the frequencies $f_0 + f_m$ and $f_0 - f_m$. Therefore, noise modeling of transistors can be split into a high and a low frequency segment. Figure 5.15 depicts the collector low-frequency noise spectral density [A^2/Hz] of a bipolar transistor.

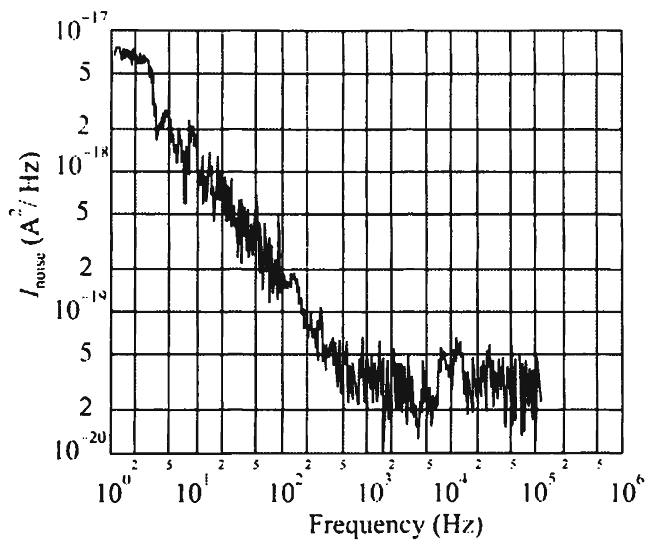


Figure 5.15 Low frequency noise spectral current density of a bipolar transistor $(A^2/\sqrt{Hz})$ measured at the collector.

5.4 Noise in Digital Signal Processing

Significantly more signal processing is done in the digital domain than the analog domain in current EW system designs. Yet communication signals are inherently analog. Thus conversion from the analog domain into the digital domain is a common and broadly implemented function. This is accomplished with *analog to digital converters* (ADCs). Likewise converting from the digital domain into the analog domain is accomplished with *digital to analog converters* (DACs).

Conversion of signals in either direction introduces additional noise into the signals being processed. The predominant conversion noise sources are discussed in this section.

Sampling a signal at a particular sample frequency f_S, has the mirroring effects illustrated in Figure 5.16. The input spectrum is alternately flipped (mirrored) and just copied to form energy regions higher in the frequency spectrum centered

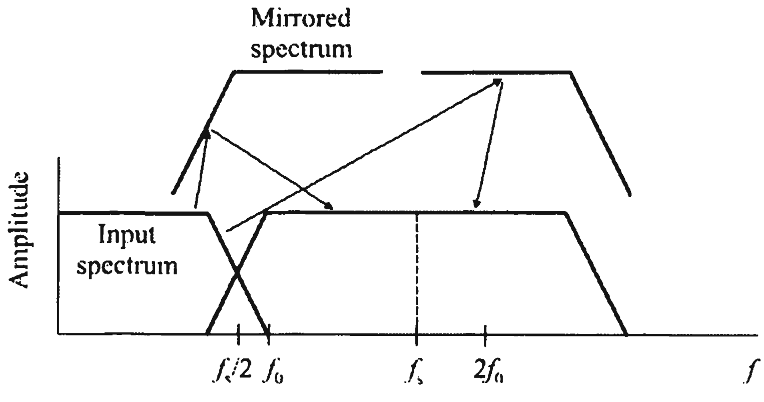


Figure 5.16 Sampling a signal mirrors and copies its spectrum at multiples of the sampling frequency.

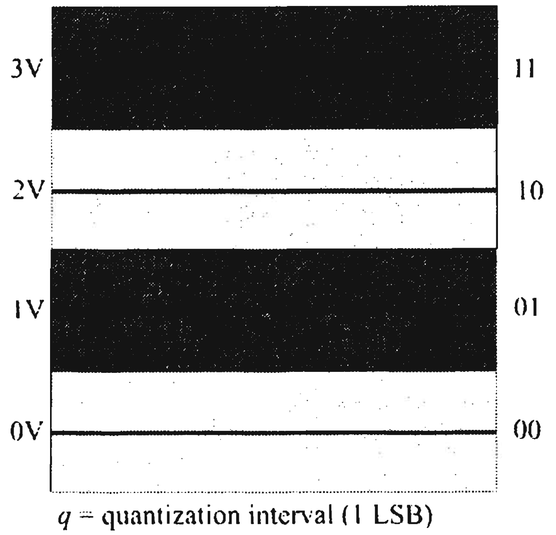


Figure 5.17 ADC quantization levels. In this case four levels are shown, corresponding to 00, 01, 10, and 11.

around multiples of f_S. We discuss aliasing shortly, but this is the root cause of it. If the sampling frequency isn't high enough to include all the energy components in the input spectrum, it is permanently distorted.

5.4.1 ADC Quantization Noise

Quantization noise is caused by quantization error. This error is actually the round-off error that occurs when an analog signal is quantized. For example, Figure 5.17 shows the output codes and corresponding input voltages for a 2-bit ADC with a 3V full-scale value. The figure shows that input values of 0V, 1V, 2V, and 3V correspond to digital output codes of 00, 01, 10, and 11, respectively. If an input of 1.75V is applied to this converter, the resulting output code would be 10 which corresponds to a 2V input. The 2V – 1.75V = 0.25V error that occurs during the quantization process is called the quantization error. Assuming the quantization error is random the quantization error can be treated as white noise.

Therefore, the quantization noise power and RMS quantization voltage for an ADC are given by

$$e_{\text{RMS}}^2 = \frac{1}{q}\int_{-q/2}^{q/2} e^2 de = \frac{q^2}{12} \qquad \text{V}^2 \tag{5.27}$$

where q is the quantization interval illustrated in Figure 5.17. So

$$e_{\mathrm{RMS}} = \frac{q}{\sqrt{12}} \tag{5.28}$$

For example, the RMS quantization noise for a 12-bit ADC with a 2.5V full-scale value is

$$e_{\mathrm{RMS}} = \frac{2.5/2^{12}}{\sqrt{12}} = 176\mu V$$

A quantized signal sampled at frequency f_s has all of its noise power folded into the frequency band of $0 \leq f \leq f_s / 2$. Assuming that this noise is random, the spectral density of the noise is given by

$$E(f) = e_{\mathrm{RMS}}\sqrt{\frac{2}{f_s}} \qquad \mathrm{V}/\sqrt{\mathrm{Hz}} \tag{5.29}$$

Converting this to noise power by squaring it and integrating over B, we get

$$n_0^2 = e_{\mathrm{RMS}}^2\left(\frac{2B}{f_s}\right) \qquad \mathrm{V}^2 \tag{5.30}$$

and

$$n_0 = e_{\mathrm{RMS}}\sqrt{\frac{2B}{f_s}} \qquad \mathrm{V} \tag{5.31}$$

where n_0 is the in-band quantization noise, B is the input signal bandwidth, and f_s is the sampling frequency. The quantity $f_s/2f_0$ is generally referred to as the *oversampling ratio* [27].

5.4.2 Data Converter Clock Jitter Noise

Clock jitter is another source of noise in the conversion process [28]. The impact is illustrated in Figure 5.18. Jitter in the time domain is equivalent to phase noise in the frequency domain. Phase noise spreads some of the clock's power away from its fundamental frequency as illustrated in Figure 5.14. This is important because sampling can be considered equivalent to mixing or multiplication in the time

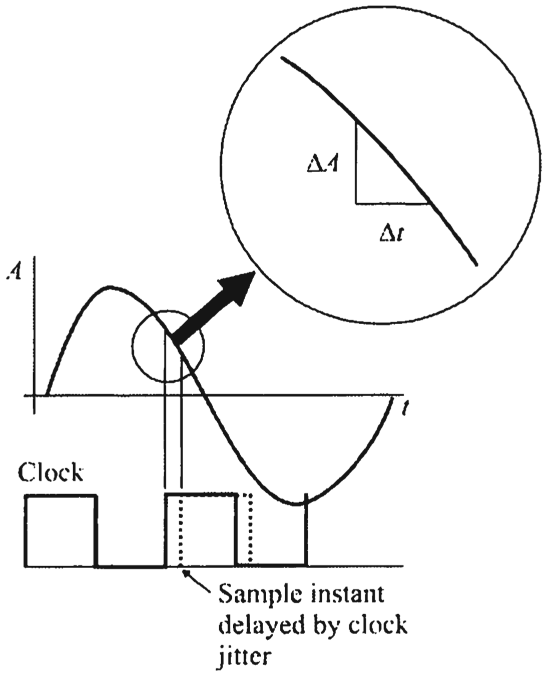


Figure 5.18 Sampling clock jitter causes sampling time error and therefore noise.

domain (the input signal gets multiplied by the sampling function, or pulse train), which is equivalent to convolution in the frequency domain. Thus, the spectrum of the sample clock is convolved with the spectrum of the input signal. Also, because jitter is wideband noise on the clock it shows up as wideband noise in the sampled spectrum. The spectrum is periodic and repeated around the sample rate. Thus, this wideband noise degrades the noise floor performance of the data converter.

A clock time delay is equivalent to a phase delay at a given frequency. In terms of noise power, this implies that phase noise in RMS radians is given by

$$\sigma_\theta^2 = \omega_{\text{clk}}^2 \sigma_t^2 \tag{5.32}$$

where σ_t is the phase jitter in RMS seconds, and ω_{clk} is the clock frequency in radians/s resulting in the effect that a higher frequency signal will have a greater phase error.

Phase noise defines the clock SNR, denoted by γ_{clk}, by

$$\gamma_{\text{clk}} = -10 \log \sigma_\theta^2 \quad \text{dB} \tag{5.33}$$

Assume the bandwidth of the clock jitter falls into a single Nyquist zone, and exclude quantization noise and thermal noise. In single carrier systems γ of a signal at f_0, sampled with a jittery clock is

$$\gamma_{\text{sig}} = \frac{1}{4\pi^2\sigma_t^2 f_0} \quad \text{dB} \tag{5.34}$$

In multiple carrier narrowband systems, γ_{sig} in dB referenced to one of the carriers, denoted as dBc, would have the same form, but the sum of all the frequency terms would replace the f_0 term in the denominator. Assuming that the power in the carriers is the same in each then

$$\gamma_{\text{sig}} = \frac{1}{4\pi^2\sigma_t^2 \sum_{i=1}^{M} f_i} \text{ dB} \tag{5.35}$$

where M is the number of carriers. This is important because it raises the quantization and thermal noise floor. So in these applications, jitter may not contribute greatly to the overall γ, and quantization and thermal noise may dominate.

In wideband systems, however, assuming the data has zero mean and a flat spectrum uniformly distributed between two frequencies f_{L} and f_{U}, it can be shown that

$$\gamma_{\text{sig}} = \frac{1}{\sigma_t^2}\frac{3}{f_{\text{U}}^2 + f_{\text{U}}f_{\text{L}} + f_{\text{L}}^2} \tag{5.36}$$

Under-sampled systems, such as those in which the clock is at a lower frequency than the signal frequency band, require clocks with much better phase jitter performance than baseband systems. This is because if the jitter is large enough, the noise caused by jitter can alias back inband (Figure 5.19). In these applications, SNR limitations due to jitter can be determined by

$$\gamma = -20\log(2\pi f_{\text{analog}} t_{\text{RMSjitter}}) \text{ dB} \tag{5.37}$$

where f_{analog} is the input frequency and $t_{\text{RMSjitter}}$ is the RMS jitter. Given a frequency of operation and an SNR requirement, the clock jitter requirement can be determined using (5.37).

Using N bits, assume a sine wave is the signal to be digitized, it has a peak-to-peak value of 2^N, an RMS value of $2^{N-1}/\sqrt{2}$, and power level of $S = 2^{2N-3}$. Since a quantization error can occur anywhere in the range of ±0.5 the pdf is given by $p(x) = 1$ over this region and the noise power is

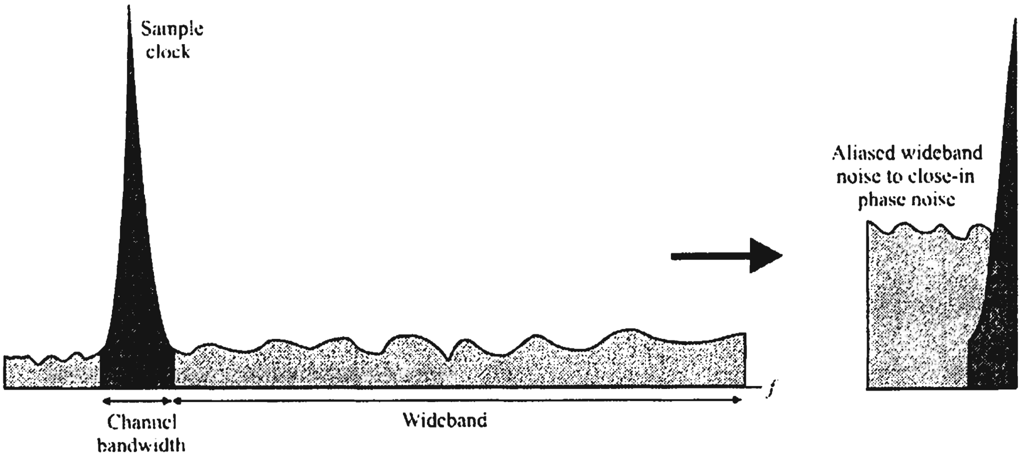


Figure 5.19 Aliased wideband noise is converted to close-in phase noise. This is the principal source of the 1/*f* shape of the phase psd around zero frequency shown in Figure 5.14.

$$N_q = \int_{-1/2}^{1/2} x^2 p(x) dx = \frac{1}{12} \tag{5.38}$$

Thus γ due to this quantization noise is given by

$$\gamma = \frac{S}{N_q} = \frac{2^{2N-3}}{1/12} = 1.5 \times 2^{2N} \tag{5.39}$$

Expressed in dB, (5.39) becomes

$$\gamma = 6N + 1.76 \qquad \text{dB} \tag{5.40}$$

which is the rule of thumb that every extra bit in digital signal processing adds an additional 6 dB to the SNR.

A first order approximation to the psd of the noise spectrum is to assume that it is flat from $f = 0$ to $f = f_s / 2$, where f_s is the sampling frequency. The actual spectrum exhibits a sin *x*/*x* shape but this approximation produces a fraction of a dB of error.

5.4.3 Sampling and Timing

All real signals exhibit phase-noise as illustrated with the psd in Figure 5.14 and timing illustrated in Figure 5.18. This noise on clocks used to trigger a sample and hold circuit associated with an ADC causes the signal to be sampled at the wrong instant yielding an incorrect value of the amplitude. The effect is illustrated in

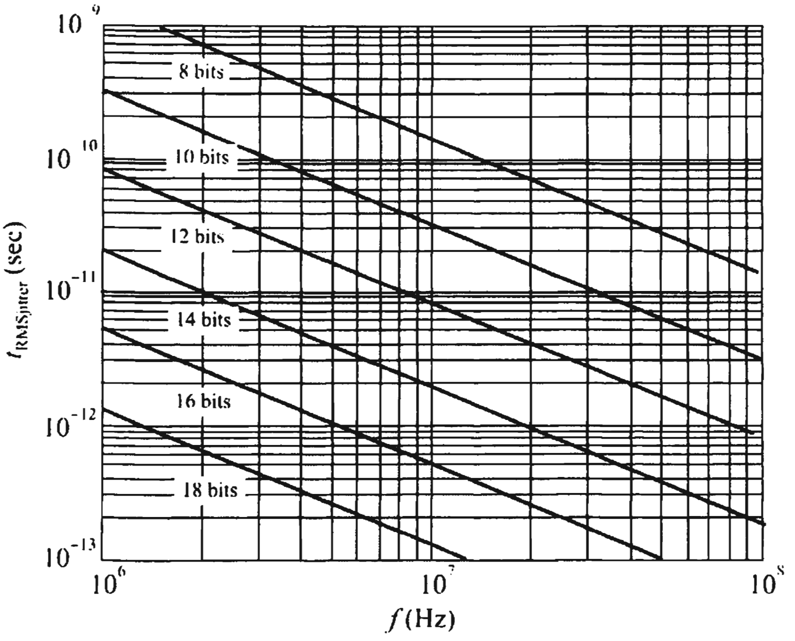


Figure 5.20 RMS jitter limits.

Figure 5.20. This sampling movement is called jitter and it reduces the accuracy of the ADC.

A figure of merit on the performance tolerance to this jitter is the amount of jitter allowed without affecting converter accuracy. This is the quantity that will be determined in the sequel.

Denote the *power spectral density* (psd) of the signal by $L(f)$. The phase jitter is determined over a range of frequencies f_1 to f_2 calculated from the psd. This phase jitter is then given by

$$\phi_j^2 = \int_{f_1}^{f_2} L(f)df \qquad \text{rad}^2 \qquad (5.41)$$

with the corresponding time jitter given by

$$t_j = \frac{\phi_j}{2\pi f_0} \qquad \text{sec} \qquad (5.42)$$

where f_0 is the center frequency, and ϕ_j is the phase jitter in radians.

Assume the signal is given by $A\sin(2\pi f_c t)$ with $A = 2^{N-1}$ sampled at f_s, and that the required jitter is to be at least four times better. Hence, the assumed required phase jitter is $\phi_j = 2^{-N}$ RMS [28]. Phase locked loop (PLL) oscillators are the modern technology of choice for providing the required accuracy and low noise jitter.

This jitter noise phase-modulates the sampling time, limiting the maximum *dV/dt* input slew rate—and, hence, maximum frequency—that will have a *dV* error less than ½ LSB as

$$f_{\max} = \frac{1}{t_{\mathrm{AJ}} 2\pi 2^{N-1}} \tag{5.43}$$

where t_{AJ} is the RMS aperture jitter and N is the nominal converter resolution. The effect of timing jitter on SNR is

$$\gamma_{\text{jitter}} = 20\log_{10} \frac{1}{2\pi f t_{\mathrm{AJ}}} \quad \text{dB} \tag{5.44}$$

Timing jitter reduces the effective SNR that can be achieved, raising the noise floor. When combining these effects, of course, they must be converted into voltage units, not added in dB.

Note that this specification assumes we are providing a jitter-free conversion clock to the converter or its *sample and hold* (S/H) circuit. The clock-source quality and careful signal routing (to minimize noise and crosstalk) are critical. For example, just 20 psec of rms jitter on a 12-bit converter with a 1 MHz input reduces the SNR by 1.5 dB. If analysis shows that the applied clock jitter may be a significant error factor, we should identify and reduce sources of jitter.

5.5 Electromagnetic Interference

Noise due to EMI is discussed in this section, which is based largely on [3, 29]. EMI (also called *radio frequency interference*, RFI) is electromagnetic radiation emitted by electrical circuits carrying rapidly changing signals which causes unwanted signals (interference or noise) to be induced in other circuits. While external man-made noise can be considered a source of EMI, in this section we consider EMI to be caused by components within the system, as man-made noise has already been discussed in Section 5.2.1. Such EMI noise can be caused and its effects felt within individual chassis in the system and/or between chassis. Cables and connectors can also be a considerable source of EMI.

The efficiency of the radiation is dependent on the height above the ground or power plane (at RF one is as good as the other) and the length of the conductor in relationship to the wavelength of the signal component (fundamental, harmonic, or transient (overshoot, undershoot, or ringing)). At lower frequencies, such as 133 MHz, radiation is almost exclusively via cables routing signals into or out of the equipment. RF noise gets onto the power planes and is coupled to the line drivers via the supply voltage and ground pins. The RF is then coupled to the cable through the line driver as common node noise. Since the noise is common mode, shielding has very little effect, even with differential pairs. The RF energy is capacitively coupled from the signal pair to the shield and the shield itself does the radiating.

The typical EMI problem is the transfer of unwanted energy from a source of interference to a circuit or system which is upset by that energy. To address the EMI problem it is necessary to determine the source of the noise, the path over which the energy is transferred, and the component that is susceptible to the interference.

5.5.1 Sources of EMI

The taxonomy of EMI emanating from sources includes separating the noise into two classes delineated by whether the interference is repetitive or consists of a single (impulse) signal.

5.5.1.1 Narrowband Repetitive Sources

Switching power supplies and computer clocks are typical of this source of EMI, which are typically repetitive pulse sources. The interference consists of a number of harmonics from the source. Note that these sources can be considered narrowband because each occurrence of a harmonic covers a narrow bandwidth. There may be many such harmonics, however, which can stretch a considerable distance in the frequency spectrum.

A repetitive pulse source can be characterized by using an amplitude versus frequency plot. The applicable variables from the time domain waveform depiction are the fundamental frequency, f_0, in Hz; the period, $T = 1/f_0$, in seconds; the peak to peak amplitude, A, in volts or amps; the pulse width, t, in seconds; pulse rise time, t_r, in seconds (faster rise times produce significant high frequency components); pulse fall time, t_f, in seconds (ditto for fall times); and the pulse duty cycle, $d = t/T$.

Using these variables, the frequency domain maximum emissions envelope can be determined. The frequency domain depiction is typically displayed as a semilog graph with the vertical axis in decibels typically relative to a microvolt or to a microampere and the horizontal axis the logarithm of frequency. Such a depiction always can always be divided into three distinct sections: the low-frequency pedestal with a 0 dB per decade slope (no slope); the middle frequency

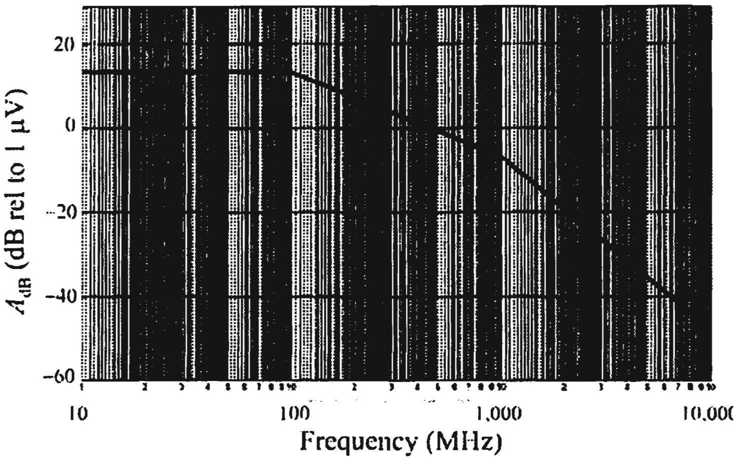


Figure 5.21 Narrowband EMI noise. All EMI sources closely match a curve such as this, although the curve breaks may be different—those shown here are representative.

section, with a –20 dB per decade slope; and the high frequency section, with a –40 dB per decade slope. Such a depiction is illustrated in Figure 5.21. The expression governing the pedestal curve is given by

$$A_{\mathrm{dB}} = 20\log_{10}(2Ad \times 10^6) \tag{5.45}$$

The first break point delineating the plateau and the middle frequency section is given by

$$f_1 = \frac{1}{\pi t} \tag{5.46}$$

Likewise the second break point delineating the middle frequency section from the high frequency section is given by

$$f_2 = \frac{1}{\pi t_{\mathrm{r}}} \tag{5.47}$$

or

$$f_2 = \frac{1}{\pi t_{\mathrm{f}}} \tag{5.48}$$

The larger of (5.47) or (5.48) is used.

The curve generated in the way just described actually represents an envelope. The amplitudes of the harmonics of the EMI noise will all be less than the magnitude of this envelope. The first harmonic will be at f_0, with higher harmonics at integral multiples of f_0. The fundamental frequency, f_0, will appear to the right of f_1, in the middle frequency section. This envelope prediction method does not give the actual amplitude at each specific harmonic frequency, as would a rigorous Fourier waveform analysis, but is much quicker than a full Fourier analysis. The answers that are provided in this way may or may not be accurate enough depending on the particular application.

5.5.1.2 Broadband Sources—Single, Impulsive Signals

When the EMI is a single pulse instead of a pulse train, the frequency domain maximum emissions envelope consists of the same three sections as described above. In this case, however, t represents the width of the single pulse. The expression governing the plateau section with 0 dB/decade slope is given by

$$A_{\mathrm{dB}} = 20\log_{10}(2At\times 10^{12}) \tag{5.49}$$

in dBμV/MHz or dBμA/MHz. The first break frequency delineating the point separating the plateau section and the middle frequency section is given by

$$f_1 = \frac{1}{\pi t} \tag{5.50}$$

and the second break point separating the middle-frequency section from the section with a –40 dB/decade slope is given by

$$f_2 = \frac{1}{\pi t_{\mathrm{r}}} \tag{5.51}$$

or

$$f_2 = \frac{1}{\pi t_{\mathrm{f}}} \tag{5.52}$$

whichever is larger.

The envelope function for a single pulse is obviously similar to that for a pulse train, with the significant difference being the units of the ordinate (dB/MHz versus dB). The other significant difference is that the energy for the single pulse starts at zero frequency and extends to infinite frequency, at least in theory, since a signal cannot simultaneously have finite time duration and limited frequency extent. There are not identifiable harmonics for the single pulse.

5.5.2 Coupling Paths

There are two types of paths over which EMI is transferred from the source to the destination: it is either conducted or radiated.

5.5.2.1 Conducted Coupling Paths

EMI can be conducted from the source to the destination over either intentional conduction paths or unintentional conduction paths, an example of the latter being the grounding of the two. A resistance is necessary in order for a current to develop a potential. There are two components to resistance in a conductor: the DC resistance and the AC resistance. The DC resistance is dictated by Ohms law. However, there is an effect in all conductors that causes the current passing through them to be carried on the outside of the wire. This skin effect forces more of the current toward the surface of the wire which increases with frequency. The decreasing cross sectional area of the conductor as the frequency increase causes the AC resistance to increase with frequency. The total resistance is the sum of the DC resistance and AC resistance.

The skin depth is the thickness at the edge of the conductor where the current is reduced to 1/*e* (~37%) of its value on the surface. It is given by [30]

$$\delta = \left(\frac{2\rho}{\omega \mu_0} \right)^{1/2}, \quad \text{in} \tag{5.53}$$

where

- ρ = Bulk resistivity of the conductor (for copper this is 6.787×10^{-7} Ω·in);
- μ_0 = Magnetic permeability of free space (= 3.192 weber/amp·in);
- ω = Frequency in radians/sec.

For example, in copper at 1 GHz, $\delta = 8.2 \times 10^{-5}$ in. Clearly, most of the RF current is conducted close to the surface of the conductor.

The amount of AC resistance generated this way is given by

$$R = (\rho \pi \mu)^{1/2} f^{1/2} \left(\frac{l}{\pi D} \right), \quad \Omega \tag{5.54}$$

with f in Hz; l, the length of the conductor under consideration (in); and D, the diameter of the conductor (in). In most cases of concern in the EW systems we are considering here, the AC resistance in wires is much greater than the DC resistance.

There are also capacitive effects when there are other conductors in the region. This capacitance can be approximated with

$$C = \frac{\varepsilon_0 A}{d}, \qquad \text{farads} \tag{5.55}$$

where

- $\varepsilon_0 = 8.84 \times 10^{-12}$ farads/m which is the dielectric constant of air;
- A = Effective plate areas of the two conductors (the smaller of the two should be used) in m^2;
- d = Distance between the two conductors, in m.

Just as in any other electrical circuit, the presence of conductor-related inductance, capacitance, and AC resistance determine the effective total impedance of conductors with respect to frequency, including resonances, and transmission line effects (i.e., characteristic impedance, proper/improper terminating impedances, energy reflections, and ringing).

5.5.2.2 Radiated Coupling Paths

There are three types of radiated coupling. The first is near-field[1] cable-to-cable coupling, also known as *crosstalk*. The second is differential mode radiation and pickup. The third is common mode radiation and pickup.

Differential mode refers to current which flows in opposite directions to two conductors, or voltage between those two conductors. Common mode refers to in-phase current flow in the supply and return conductors, or voltage from either of the two conductors to local ground.

Crosstalk

Near field coupling from one cable to another is driven by electric fields and/or magnetic fields emanating from the source cable and impinging upon the receptor cable. The efficiency of signal transfer from the source cable to the receptor cable

[1] Since this is the near field, the signal is not really radiated, but coupled nevertheless.

depends upon cable spacing, respective cable orientations, frequency, and termination impedances. Crosstalk can be minimized by increasing the separation distances between the cables, or by crossing the cables at right angles. A general fix for crosstalk is not including dissimilar categories of conductors in the same cable bundle and especially keeping digital lines as far away from sensitive RF inputs as possible.

Differential Mode Radiation

Circuits act like radiating loop antennas. If the physical dimensions of a circuit are short compared to a wavelength at the circuit signal frequency or harmonics of concern, the electric signal strength of differential mode radiation is given by

$$E_{\mathrm{DM}} = 263 \times 10^{-16} \frac{f^2 A I}{r}, \quad \mathrm{V/m} \tag{5.56}$$

where

- f = Frequency in Hz;
- A = Circuit loop area in m^2;
- I = Current in amperes; and
- r = Distance between radiating circuit and affected conductor, in m.

Note that the radiation is proportional to the square of the driving frequency which imposes severe penalties for increasing digital clock speeds.

Differential Mode Pickup

Circuits also act like receiving loop antennas, at least up to the frequency at which the longest dimension of the physical circuit loop represents a half wavelength at the incoming interference frequency in question. The circuit voltage pickup can be approximated with

$$V = ELWf \frac{\pi}{150}, \quad \mathrm{V} \tag{5.57}$$

where

- E = Incident field strength in V/m;
- L = Length of circuit loop in meters (must be less than $\lambda/2$);

- W = Width of circuit loop in meters; and
- f = Frequency of incident field, in MHz.

For $L > \lambda/2$,

$$V = \pi E W, \text{ V} \tag{5.58}$$

Common Mode Radiation

Common mode radiation can radiate from a loop conductor carrying common mode current, in which case (5.56) can be used. In many cases, however, only a very high impedance "loop" exists, such as that of a single long conductor, with attached electronic circuitry, with the loop closed by a small parasitic capacitance to local ground. In this case, the radiation is modeled as coming from a grounded monopole conductor, as approximated by

$$E_{\text{CM}} = 12.6 \times 10^{-7} \frac{f\,I\,L}{r} \tag{5.59}$$

where

- f = Frequency in Hz;
- I = Common mode current in amperes;
- L = Length of the common mode circuit/cable combination in m; and
- r = Distance from radiating common mode circuit, in m.

Common Mode Pickup

A common mode receiving loop will pick up energy per (5.57). As can be seen from the above equations, radiation and pickup are both affected by frequency and circuit dimensions. What is not shown in the equations is a dependence upon circuit and field orientation. This omission is due to the simplification of the relevant equations and apply to the worst case.

5.5.2.3 EMI Shielding

The isolation of circuits from the environment by means of a conductive barrier is called *shielding*. Shielding is almost always required in components in EW systems because of the sensitivity of the receivers. For low-frequency magnetic fields, high permeability shields are necessary. Shielding performance is

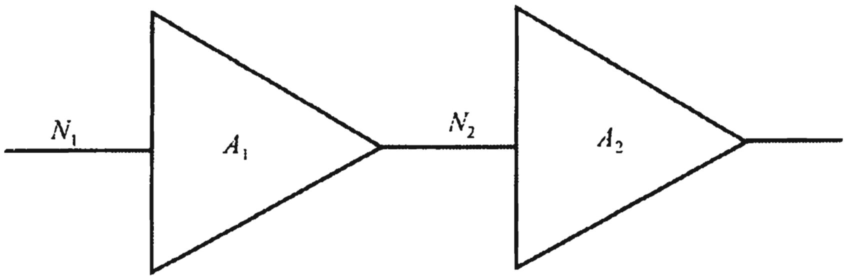


Figure 5.22 Cascaded amplifiers. N_i corresponds to the noise power at the nodes shown.

compromised by any form of discontinuity such as penetrating wires and apertures.

5.5.3 Remarks

Effective treatment of EMI problems depends upon thoroughly understanding the parts of the problem discussed here (source, path, and victim), the degree of the problem, and the required level of improvement. Probably the best solution is to design to minimize EMI in the first place.

Communication EW systems, of necessity, have very sensitive front-ends. If EMI signals with significant amplitudes are present, whatever the cause, these sensitive circuits can quickly become overloaded. This is called *desensitization*. The receiver chain typically has a dynamic range of about 70 dB, so if a local EMI signal is greater than the noise floor plus this dynamic range, the entire receiving chain becomes desensitized.

5.6 Noise in Amplifiers

We will investigate the effects of noisy amplifiers in this section [31–37]. Consider a chain of two amplifiers (or amplifying devices), with power gains A_1 and A_2, and input noise power levels N_1 and N_2 as illustrated in Figure 5.22. A signal with power S is applied to the first amplifier, so the input signal-to-noise ratio is S/N_1. At the output of the first amplifier the signal power is A_1S and the noise power is A_1N_1. Both are amplified by the second amplifier, but in addition the second amplifier contributes its noise, so the signal-to-noise ratio at the output of the second amplifier is

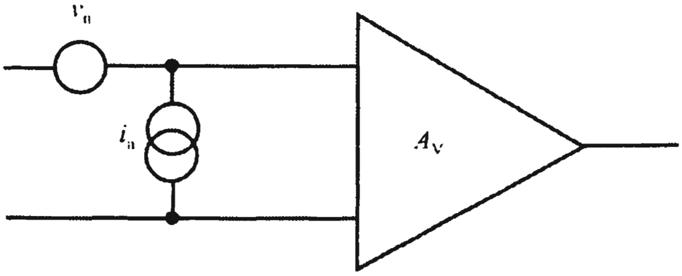


Figure 5.23 Amplifier noise model. There are both a voltage noise source and current noise source at the input. The resulting amplifier (A_V) is noise-free.

$$\left(\frac{S}{N}\right)^2 = \frac{(SA_1A_2)^2}{(N_1A_1A_2)^2+(N_2A_2)^2} = \frac{S^2}{N_1^2+\left(\frac{N_2}{A_1}\right)^2}$$

$$= \left(\frac{S}{N_1}\right)^2 \frac{1}{1+\left(\frac{N_2}{A_1N_1}\right)^2} \tag{5.60}$$

The overall signal-to-noise ratio is reduced, but the noise contribution from the second stage can be negligible, provided the gain of the first stage is sufficiently high. Therefore, in a well-designed system the noise is dominated by the first gain stage.

5.6.1 Amplifier Noise Model

The noise properties of any amplifier can be described fully in terms of a voltage noise source and current noise source at the amplifier input, as illustrated in Figure 5.23 [38]. Typical magnitudes are in the nanoVolts/Hz (nV/Hz) and femtoAmperes/Hz (fA/Hz) to picoAmperes/Hz (pA/ Hz). Here the magnitude of the noise sources is characterized by the spectral density. The noise sources do not have to physically be present at the input. Noise also originates within the amplifier. Assume that at the output the combined contribution of all internal noise sources has the spectral density e_{no}. If the amplifier has a voltage gain A_V , this is equivalent to a voltage noise source at the input $e_n = e_{no}/A_V$. It is convenient to express the input noise in terms of spectral density, so that the effect of amplifier bandwidth can be assessed separately.

Assume that a sensor with resistance R_S is connected to an amplifier with voltage gain A_V and an infinite input resistance, so no current flows into the amplifier, as illustrated in Figure 5.24. The input noise current i_n flows through the source resistance R_S to yield a noise voltage i_nR_S, which adds to the thermal noise

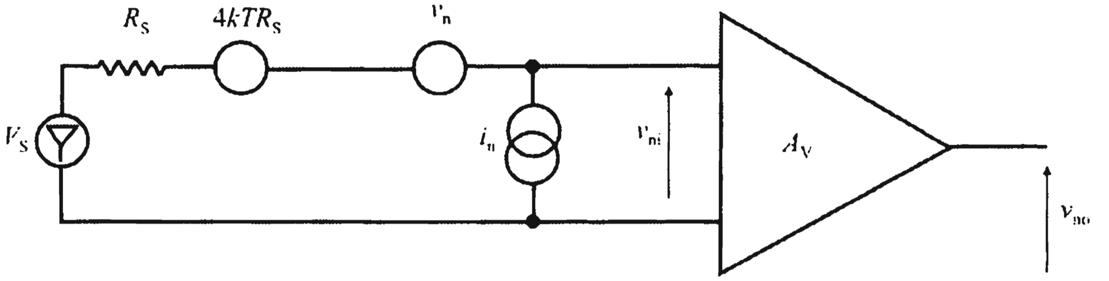


Figure 5.24 Amplifier driven by a noisy antenna source. The antenna noise voltage is represented by the $4kTR_S$ voltage source. The resulting antenna and R_S are noise free.

of the source resistance and the noise voltage of the amplifier. All terms add in quadrature, since they are not correlated. The total noise voltage at the input of the amplifier is

$$v_{ni}^2 = 4kTR_S + v_n^2 + (i_n R_S)^2 \tag{5.61}$$

and at the output of the amplifier

$$v_{no}^2 = (A_V v_{ni})^2 = A_V^2[4kTR_S + v_n^2 + (i_n R_S)^2] \tag{5.62}$$

The signal-to-noise ratio at the amplifier output is

$$\left(\frac{S}{N}\right)^2 = \frac{A_V^2 V_S^2}{A_V^2[4kTR_S + v_n^2 + (i_n R_S)^2]} \tag{5.63}$$

and is independent of the amplifier gain and equal to the input SNR, as both the input noise and the signal are amplified by the same amount.

In the preceding example the amplifier had an infinite input resistance, so no current flowed into the amplifier. Assume now that the input impedance is finite as illustrated in Figure 5.25. The signal at the input of the amplifier is

$$V_{S_i} = V_S \frac{R_i}{R_S + R_i} \tag{5.64}$$

The noise voltage at the input of the amplifier is expressed as

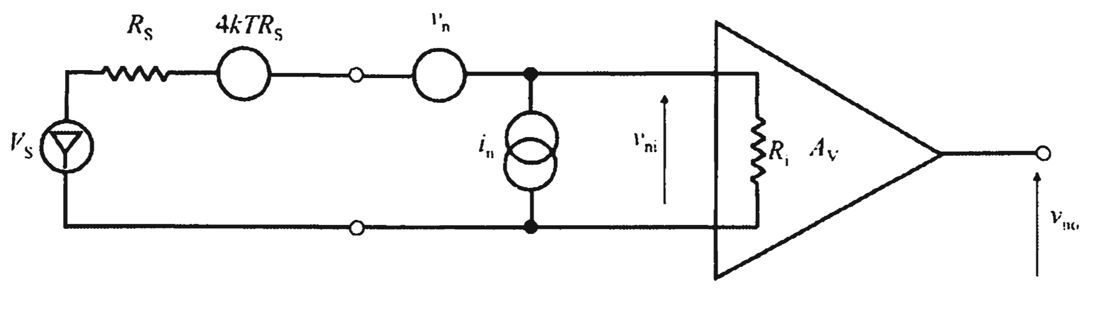


Figure 5.25 Amplifier with input resistance, R_i.

$$v_{n_i}^2 = [4kTR_S + v_n^2]\left(\frac{R_i}{R_i + R_S}\right)^2 + i_n^2\left(\frac{R_i R_S}{R_i + R_S}\right)^2 \tag{5.65}$$

where the bracket in the i_n^2 represents the parallel combination of R_i and R$_S$. The signal-to-noise ratio at the output of the amplifier

$$\left(\frac{S}{N}\right)_o^2 = \frac{A_V^2 V_{S_i}^2}{A_V^2 v_{n_i}^2} = \frac{V_S^2\left(\frac{R_i}{R_S + R_i}\right)^2}{(4kTR_S + v_n^2)\left(\frac{R_i}{R_S + R_i}\right)^2 + i_n^2\left(\frac{R_i R_S}{R_S + R_i}\right)^2}$$

$$= \frac{V_S^2}{(4kTR_S + v_n^2) + i_n^2 R_S^2} \tag{5.66}$$

is the same as for an infinite input resistance. Therefore the SNR$_o$ *is independent of amplifier input impedance.*

Even though this derivation used resistances, the result obviously also holds for a complex input impedance, i.e., a combination of resistive and capacitive or inductive components.

The noise sources can be correlated with correlation coefficient γ. Then

$$v_n^2 = v_{n_1}^2 + v_{n_2}^2 + 2\gamma v_{n_1} v_{n_2} \tag{5.67}$$

in the above example; if the input noise voltage and current are correlated, the input noise voltage becomes

$$v_{ni}^2 = 4kTR_S + v_n^2 + i_n^2 R_S^2 + 2\gamma v_n i_n R_S \tag{5.68}$$

The total noise at the output is obtained by integrating over the spectral noise power

$$P_{\mathrm{n}}(f) \propto v_{\mathrm{no}}^2(f) \tag{5.69}$$

The noise level at the output is determined by the frequency distribution of the noise at the output, as given by the noise psd $e_{\mathrm{no}}(f)$ which, in turn, is determined both by the spectral distribution of the input noise voltage and current and by the frequency response of the amplifier. Thus

$$v_{\mathrm{n_o}}^2 = \int_0^\infty e_{\mathrm{n_o}}^2(f)df = \int_0^\infty e_{\mathrm{n_i}}^2(f)\left|A_{\mathrm{V}}(f)\right|^2 df \tag{5.70}$$

The amplifier gain factor is shown as magnitude squared, because normally the amplifier has a frequency dependent gain and phase characteristic, so it is complex.

5.6.2 Noise Bandwidth Versus Signal Bandwidth

Consider an amplifier with the frequency response $A(f) = A_0 G(f)$, where A_0 is the maximum gain and $G(f)$ describes the frequency response. For example, for the simple amplifier described above

$$A_{\mathrm{v}} = g_{\mathrm{m}}\left(\frac{1}{R_{\mathrm{L}}} + j\omega C_0\right)^{-1} = g_{\mathrm{m}} R_{\mathrm{L}} \frac{1}{1 + j\omega R_{\mathrm{L}} C_0} \tag{5.71}$$

so that

$$A_0 = g_{\mathrm{m}} R_{\mathrm{L}} \tag{5.72}$$

and

$$G(f) = \frac{1}{1 + j\omega R_{\mathrm{L}} C_0} \tag{5.73}$$

If a white noise source (flat spectral density) with spectral density e_{ni} is present at the input, the total noise voltage at the output is

$$v_{\mathrm{n_o}} = \sqrt{\int_0^\infty e_{\mathrm{n_i}}^2 \left|A_0 G(f)\right|^2 df} = e_{\mathrm{n_i}} A_0 \sqrt{\int_0^\infty G^2(f) df} = e_{\mathrm{n_i}} A_0 \sqrt{\Delta f_{\mathrm{n}}} \tag{5.74}$$

Δf_{n} is the called the *noise bandwidth.*

In general, the noise bandwidth and the signal bandwidth are not the same. If the upper cutoff frequency is determined by a single *RC* time constant, as in the Lorentzian, the signal bandwidth is

$$\Delta f_{\mathrm{s}} = f_{\mathrm{U}} = \frac{1}{2\pi RC} \tag{5.75}$$

and the noise bandwidth is

$$\Delta f_{\mathrm{n}} = \frac{1}{4RC} = \frac{\pi}{2} f_{\mathrm{u}} \tag{5.76}$$

5.6.3 Noise Bandwidth and Low-Frequency (1/*f*) Noise

For the low frequency noise psd given by

$$P_{\mathrm{nf}} = \frac{S_{\mathrm{f}}}{f} \tag{5.77}$$

and the corresponding voltage density

$$e_{\mathrm{nf}}^2 = \frac{A_{\mathrm{f}}}{f} \tag{5.78}$$

the total noise power in a frequency band f_1 to f_2 is given by the integral of (5.78)

$$v_{\mathrm{nf}}^2 = \int_{f_1}^{f_2} \frac{A_{\mathrm{f}}}{f} df = A_{\mathrm{f}} \ln\left(\frac{f_2}{f_1}\right) \tag{5.79}$$

Thus, for a 1/*f* spectrum the total noise power depends on the ratio of the upper to lower cutoff frequency.

Frequently, the 1/*f* noise comer is specified which is that frequency where 1/*f* noise intercepts the white noise power level. Higher white noise levels reduce the comer frequency, so a lower noise comer does not equate to lower 1/*f* noise.

5.6.4 Noise Factor and Noise Figure

A key parameter of electronic system analysis is the SNR given by

$$\gamma = \text{SNR} = \frac{S}{N} = \frac{\text{Signal Power}}{\text{Noise Power}} \tag{5.80}$$

A high SNR means that the signal can be identified, whereas a low SNR means that the signal is obscured by noise. The noise factor is a measure of the degradation in SNR as a signal passes through any processing stage. The definition of *noise factor* (f_a) is the ratio of the total available noise power at the stage output to the available noise power at the output due to the input noise alone

$$f_\text{a} = \frac{\text{Output Noise Power}}{\text{Ideal Output Noise Power} = \text{Gain}\times\text{Input Noise Power}} \geq 1 \tag{5.81}$$

When the noise factor is expressed in dB, it is called the *noise figure* and is given by

$$F_\text{a} = 10\log_{10} f_\text{a} \tag{5.82}$$

For an ideal component which adds no noise to the output noise power, the noise is all due to the input noise power so $f_\text{a} = 1$ ($F_\text{a} = 0$ dB). The gain used in this expression is the available power gain

$$G_\text{A} = \frac{P_\text{out}}{P_\text{in}} = \frac{S_\text{o}}{S_\text{i}} \tag{5.83}$$

We can easily see that the noise factor is equal to the ratio of the input SNR to the output SNR:

$$f_\text{a} = \frac{N_\text{o}}{N_\text{i} G_\text{A}} = \frac{N_\text{o}}{N_\text{i} S_\text{o} / S_\text{i}} = \frac{S_\text{i}/N_\text{i}}{S_\text{o}/N_\text{o}} \tag{5.84}$$

We can also write

$$f_{\mathrm{a}} = \frac{G_{\mathrm{A}} N_{\mathrm{i}} + P_{\mathrm{n}}}{N_{\mathrm{i}} G_{\mathrm{A}}} = 1 + \frac{P_{\mathrm{n}}}{G_{\mathrm{A}} N_{\mathrm{i}}} \tag{5.85}$$

Since noise factor is a dimensionless quantity, it is often expressed as the noise figure.

At an operating temperature of *T* degrees, the noise factor can also be written as

$$f_{\mathrm{a}} = 1 + \frac{1 - G_{\mathrm{A}}}{G_{\mathrm{A}}} \frac{T}{T_0} \tag{5.86}$$

where the reference temperature T_0 = 290 K (approximately room temperature). When the amplifier is at room temperature, the noise factor is therefore equal to the power loss factor of the network, L_{A}:

$$f_{\mathrm{a}} = \frac{1}{G_{\mathrm{A}}} = L_{\mathrm{A}} \tag{5.87}$$

5.6.4.1 Lossy Components

The noise figure of a lossy component can be determined by recognizing that the noise power at the output of the component must be the same as the noise power at the input. (If that were not the case then power would flow in the component due to the difference in power — thermal equilibrium.) Thus

$$G_{\mathrm{A}} N_{\mathrm{i}} + G_{\mathrm{A}} N_{\mathrm{added}} = N_{\mathrm{i}} \tag{5.88}$$

Solving for the equivalent additional power at the input ($G_{\mathrm{A}}N_{\mathrm{added}}$) gives $G_{\mathrm{A}} N_{\mathrm{added}} = N_{\mathrm{i}}(1 - G_{\mathrm{A}})$. The noise factor is then

$$f_{\mathrm{a}} = 1 + \frac{\text{Noise Added}}{G_{\mathrm{A}} N_{\mathrm{i}}} = 1 + \frac{G_{\mathrm{A}} N_{\mathrm{added}}}{G_{\mathrm{A}} N_{\mathrm{I}}} = 1 + \frac{N_{\mathrm{i}}(1 - G_{\mathrm{A}})}{G_{\mathrm{A}} N_{\mathrm{i}}} = \frac{1}{G_{\mathrm{A}}} = L \tag{5.89}$$

where L is the power loss of the device. Thus, the noise factor is the same as the loss. A lossy system component such as a length of lossy transmission line, therefore, leads to degradation in SNR.

5.6.4.2 Cascaded Networks

Suppose we have two stages in a system,

$$N_\text{o} = G_{\text{A}_2} N_{\text{o}_1} + P_{\text{n}_2} = G_{\text{A}_2}(G_{\text{A}_1} N_\text{i} + P_{\text{n}_1}) + P_{\text{n}_2} \tag{5.90}$$

$$f_\text{a} = \frac{G_{\text{A}_2}(G_{\text{A}_1} N_\text{i} + P_{\text{n}_1}) + P_{\text{n}_2}}{N_\text{i} G_{\text{A}_1} G_{\text{A}_2}} = 1 + \frac{P_{\text{n}_1}}{N_\text{i} G_{\text{A}_1}} + \frac{P_{\text{n}_2}}{N_\text{i} G_{\text{A}_1} G_{\text{A}_2}} \tag{5.91}$$

If $G_{\text{A}1}$ is large, the first stage is dominant in obtaining a low factor because $G_{\text{A}1}$ appears in the denominator of the second and third terms (and, in fact, all the remaining terms when there are more than two stages).

In terms of the noise factors of the two stages,

$$f_{\text{a}_1} = 1 + \frac{P_{\text{n}_1}}{N_\text{i} G_{\text{A}_1}} \tag{5.92}$$

$$f_{\text{a}_2} = 1 + \frac{P_{\text{n}_2}}{N_\text{i} G_{\text{A}_2}} \tag{5.93}$$

the noise factor of the system is

$$f_\text{a} = f_{\text{a}_1} + \frac{f_{\text{a}_2} - 1}{G_{\text{A}_1}} \tag{5.94}$$

The noise factor contribution of the second stage is divided by the gain of the first stage, thereby reducing its influence. The amount of reduction, of course, depends on the gain of the first stage. Again, we can see that the first stage is most critical in determining the noise factor of the system, and for noise reduction purposes, should have as much gain as possible.

When n is noisy, linear stages are cascaded, which is usually the case, then the resultant noise factor is given by Friis's formula [39]

$$f_\text{a} = f_{\text{a}_1} + \frac{f_{\text{a}_2} - 1}{G_1} + \frac{f_{\text{a}_3} - 1}{G_1 G_2} + \cdots + \frac{f_{\text{a}_n} - 1}{G_1 G_2 \ldots G_{n-1}} \tag{5.95}$$

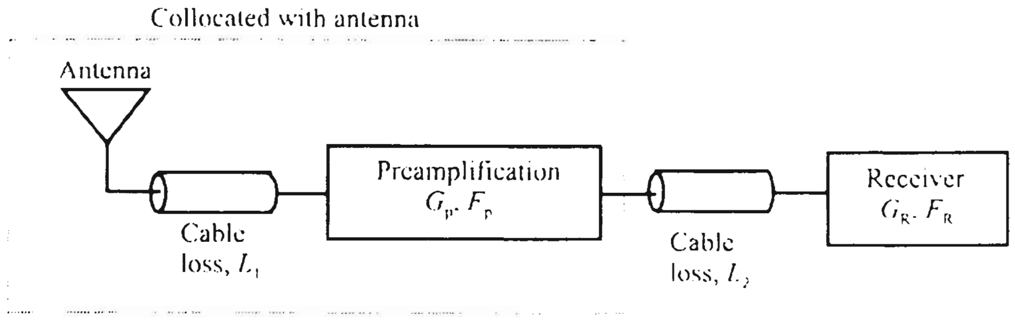


Figure 5.26 Components in an RF chain in an EW system prior to the receiver.

where the gains of the individual stages are given by G_i and their noise factors are given by f_i. In a typical communication EW receiver, the linear stages to which this noise figure calculation applies consist of the RF amplification stages as well as the IF amplifier stages—everything before the demodulation of the signals. Therefore, to minimize the noise figure of the system, it is important to use high gain stages early in the RF path.

When a receiver is embedded into a system, then the system noise factor is defined in an equivalent fashion, but cable losses, signal distribution losses, preamplification gains, and the like must be taken into consideration. The components in an RF receiver chain prior to the receiver are shown in Figure 5.26. They include antennas, cables, and perhaps some amplification at the antenna. Losses prior to the first amplification stage are simply added to the noise factor of that stage. With all parameters expressed in decibels, the system noise figure can be computed by using the following equation

$$F_a\,(\text{in dB}) = 10\log\left[\begin{matrix}10^{F_{ap}/10} + 10^{L_1/10} + 10^{(L_2-G_p)/10} \\ +10^{(F_{aR}-G_p)/10} - 10^{-G_p/10}\end{matrix}\right] \tag{5.96}$$

F_{aR} is the receiver noise figure, F_{ap} is the noise figure of the preamplifier, and L_2 is the loss between the preamplifier and the receiver. This equation assumes that the gain of the receiver is sufficiently high that all stages following the receiver contribute negligibly to the system noise figure.

5.6.4.3 Noise Matching

The noise factor can be expressed as

$$f_{\text{a}} = \frac{v_{\text{n}_{\text{i}}}^2}{4kTR_{\text{S}}} = 1 + \frac{v_{\text{n}}^2 + (i_{\text{n}}R_{\text{S}})^2}{4kTR_{\text{S}}} = 1 + \frac{v_{\text{n}}^2}{4kTR_{\text{S}}} + \frac{i_{\text{n}}^2 R_{\text{S}}}{4kT} \tag{5.97}$$

Thus the noise factor assumes a minimum when

$$R_{\text{S}} = \frac{v_{\text{n}}}{i_{\text{n}}} \tag{5.98}$$

In a matched system with a resistive source

$$F_{\text{a,opt}} = \frac{v_{\text{n}} i_{\text{n}}}{2kT} \tag{5.99}$$

This principle of "noise matching" must be applied with caution, however:

1. Power is not always the relevant measure. Sometimes the noise voltage is most important. Minimum noise voltage e_{ni} always results when $R_{\text{S}} = 0$.

2. Merely increasing the source resistance will increase the total input noise v_{ni} without improving the SNR. The advantage of noise matching only results when both the signal and the effective source resistance are modified simultaneously, for example, by a transformer.

Noise Matching with a Transformer

One way of achieving satisfaction of (5.98) is to insert a transformer between the antenna and the amplifier as shown in Figure 5.27. The antenna is coupled to the amplifier through a transformer with the turns ratio $N = N_{\text{S}}/N_{\text{P}}$. Assume unity coupling in the transformer. Then the signal appearing at the secondary is

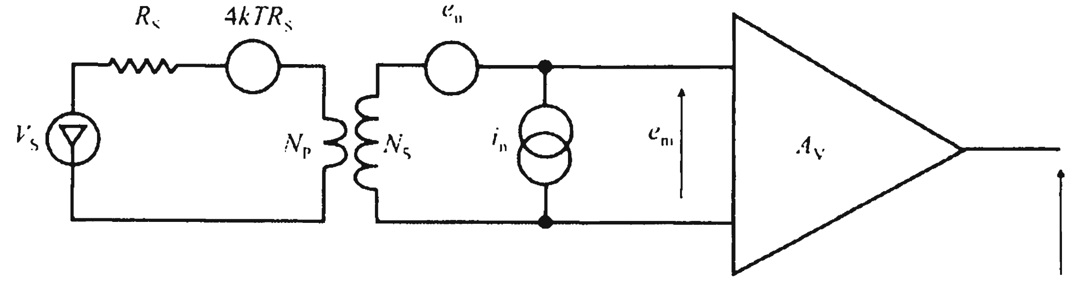


Figure 5.27 Transformer coupled antenna/amplifier.

$$V_{ss} = NV_S \tag{5.100}$$

The thermal noise of the source at the secondary is given by

$$v_{nSS}^2 = N^2 4kTR_S \tag{5.101}$$

Because the transformer also converts impedances, the source resistance appears at the secondary as

$$R_{SS} = N^2 R_S \tag{5.102}$$

Thus, the signal is increased, but so is the noise contribution due to the input noise current

$$v_{n_i}^2 = 4kTR_S N^2 + v_n^2 + R_S^2 N^4 i_n^2 \tag{5.103}$$

and the signal-to-noise ratio

$$\gamma^2 = \frac{V_S^2 N^2}{4kTR_S N^2 + v_n^2 + R_S^2 N^4 i_n^2} = \frac{V_S^2}{4kTR_S + \frac{v_n^2}{N^2} + R_S^2 N^2 i_n^2} \tag{5.104}$$

that attains a maximum for

$$R_S N^2 = \frac{v_n}{i_n} \tag{5.105}$$

Signal-to-Noise Ratio Maximization

Consider Figure 5.28. The goal is to find R_i and N_s to maximize γ out of the amplifier. It is assumed that the reactive parts of the source and input impedance have been matched. In this case the transformer is the matching element. It is assumed that the transformer is noiseless as is the amplifier. The power of the signal at the output is given by

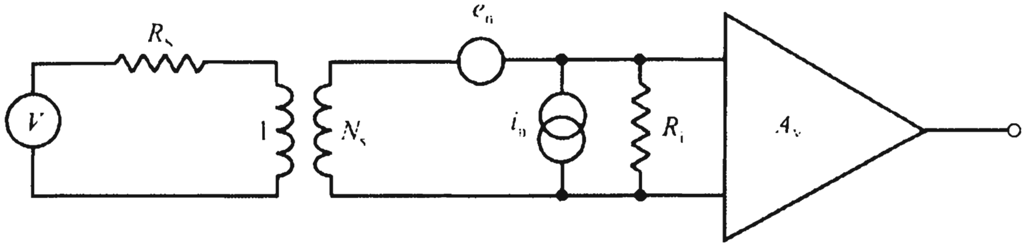


Figure 5.28 Maximizing the SNR by varying R_i and the turns ratio $1{:}N_s$.

$$P_s = \frac{V^2 N_s^2}{\left[G_i\left(\frac{1}{G_i} + \frac{N_s^2}{G_s}\right)\right]^2} A_v^2 \tag{5.106}$$

where $G_i = 1/R_i$ and $G_S = 1/R_S$. Likewise the noise power at the output is given by

$$P_n = e_n^2 k^2 + \frac{i_n^2}{\left(G_i + \frac{G_s}{N_s^2}\right)^2} A_v^2 \tag{5.107}$$

where it is assumed that the noise voltage, noise current, and the signal are all independent.

γ at the output is therefore given by

$$\gamma = \frac{\dfrac{V^2 N_s^2}{\left(1 + N_s^2 \dfrac{G_i}{G_s}\right)^2}}{e_n^2 + \dfrac{i_n^2}{\left(G_i + \dfrac{G_s}{N_s^2}\right)^2}} \tag{5.108}$$

The maximum SNR is found by taking the derivative and setting it equal to zero yielding

$$N_{\mathrm{s}}^{4}=\frac{G_{\mathrm{s}}^{2}}{G_{\mathrm{i}}^{2}+{i_{\mathrm{n}}^{2}}/{e_{\mathrm{n}}^{2}}} \tag{5.109}$$

When $R_{\mathrm{i}}\rightarrow\infty$ ($G_{\mathrm{i}}\rightarrow 0$) we get

$$N_{\mathrm{s}}^{2}=G_{\mathrm{s}}\frac{e_{\mathrm{n}}}{i_{\mathrm{n}}} \tag{5.110}$$

By choosing $N_{\mathrm{s}} = 1$, we get $R_{\mathrm{S}} = e_{\mathrm{n}}/i_{\mathrm{n}}$, which is the noise resistance of the amplifier. Alternately, we can first set $N_{\mathrm{s}} = 1$ and solve for R_{S}.

$$R_{\mathrm{S}}=R_{\mathrm{i}}\sqrt{\frac{\left({e_{\mathrm{n}}}/{i_{\mathrm{n}}}\right)^{2}}{R_{\mathrm{i}}^{2}+\left({e_{\mathrm{n}}}/{i_{\mathrm{n}}}\right)^{2}}} \tag{5.111}$$

Depending on the values of R_{i} and $e_{\mathrm{n}}/i_{\mathrm{n}}$, we get two different solutions. The first one is noise matching and the second one is power matching.

$$R_{\mathrm{S}}\gg\frac{e_{\mathrm{n}}}{i_{\mathrm{n}}},\qquad R_{\mathrm{i}}\gg\frac{e_{\mathrm{n}}}{i_{\mathrm{n}}} \tag{5.112}$$

and

$$R_{\mathrm{s}}\approx R_{\mathrm{i}},\qquad R_{\mathrm{i}}\ll\frac{e_{\mathrm{n}}}{i_{\mathrm{n}}} \tag{5.113}$$

By substituting (5.109) into (5.108) we get

$$\gamma=\frac{V^{2}}{2R_{\mathrm{s}}e_{\mathrm{n}}^{2}\left(\frac{1}{R_{\mathrm{i}}}+\sqrt{\frac{1}{R_{\mathrm{i}}^{2}}+\frac{i_{\mathrm{n}}^{2}}{e_{\mathrm{n}}^{2}}}\right)} \tag{5.114}$$

Let $R_{\mathrm{i}}=\infty$ and find the values of i_{n} and e_{n} in order to observe the thermal noise of R_{S} with $\gamma = 1$. The signal V is therefore the thermal noise of R_{S}.

$$V^2 = 4R_\mathrm{S}k_\mathrm{B}T = 2R_\mathrm{S}i_\mathrm{n}e_\mathrm{n} \qquad (5.115)$$

where T is the temperature in Kelvins and k_B is Boltzman's constant. We can now find the required temperature of R_S. Because we have chosen $\gamma = 1$, that temperature is the "equivalent" temperature of the amplifier.

$$T = \frac{i_\mathrm{n}e_\mathrm{n}}{2k_\mathrm{B}} \qquad (5.116)$$

5.6.5 Equivalent Noise Temperature

We can also express the "noisiness" of a component in terms of an equivalent noise temperature using $P = k_\mathrm{B}TB$

$$f_\mathrm{a} = \frac{S_\mathrm{i} / k_\mathrm{B}T_0B}{G_\mathrm{A}S_\mathrm{i} / G_\mathrm{A}k_\mathrm{B}(T_0 + T_\mathrm{e})B} = 1 + \frac{T_\mathrm{e}}{T_0} \qquad (5.117)$$

[Note that this T_e is not the same T in (5.86).] When specifying the equivalent temperature T_e of a component, we assume that the input noise power corresponds to room temperature, so that $T_0 = 290$ K. Equivalent temperature is most useful for low noise figure devices.

5.7 Other Noise Considerations

5.7.1 Correlated Noise

The noise sources in communication EW systems are generally independent of each other so the total noise power is additive

$$P_\mathrm{n,total} = \sum_{i=1}^{M} P_{\mathrm{n},i} \qquad (5.118)$$

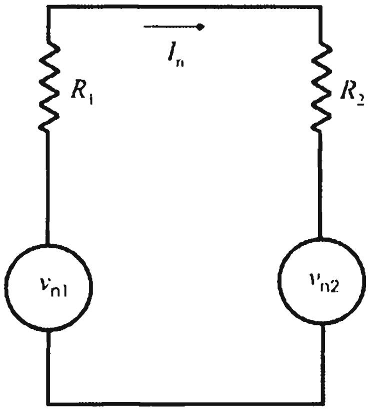


Figure 5.29 Correlated noise current.

for M independent noise sources. However, in some cases two noise parameters are coherent. Consider Figure 5.29, for example, where the noise current through the two resistors is fully correlated since, of necessity, it is equal. Coherent systems are sensitive to relative phase. For two correlated noise sources N_1 and N_2 the total noise is

$$N^2 = N_1^2 + N_2^2 + 2\gamma N_1 N_2 \qquad (5.119)$$

where the correlation coefficient, γ, ranges from −1 (anticorrelated, i.e., identical, but 180° out of phase) to +1 (identical, fully correlated). For uncorrelated noise components $\gamma = 0$ and then individual current or voltage noise contributions add in quadrature, e.g.,

$$v_{\text{n,total}} = \sqrt{\sum_{i=1}^{M} v_{\text{n},i}^2} \qquad (5.120)$$

or

$$i_{\text{n,total}} = \sqrt{\sum_{i=1}^{M} i_{\text{n},i}^2} \qquad (5.121)$$

5.7.2 Partition Noise

Partition noise in a *bipolar junction transistor* (BJT) is due to recombination at the junction, which produces a noise current of

$$i_\mathrm{p}^2 = \frac{2kT}{r_\mathrm{e}'}\alpha_0(1-\alpha_0)B \tag{5.122}$$

where $kT = -174$ dBm in a 1 Hz bandwidth, B is the bandwidth, r_e' is the emitter resistance, and α_0 is the current gain of the BJT.

5.8 Concluding Remarks

The predominant sources of noise that affect communication EW systems, both external and internal, were presented in this chapter.

The principal external sources are atmospheric, thermal, and man-made. Atmospheric noise is predominant in the lower RF range, while thermal noise is constant pretty much everywhere in the RF spectrum of interest here. Man-made noise, by its very nature, is fairly unpredictable. In the environment that EW systems find themselves—situations that involve many sources of man-made noise—this can be the dominant source of external noise.

Several internal sources were also discussed. Again, thermal noise is an internal source, caused by any electronic device that has a resistive component—and that is all of them. Active electronic devices all have additional noise sources that are unique to the technology from which they are made. EMI is a noise source that is caused by unintended radiation coupling from one device to another. In sensitive RF systems such as EW systems, EMI can be a major problem and EMI protection must be designed in from the start if it is to be controlled.

Noise in amplifiers was discussed at length since amplification is a ubiquitous requirement in almost all circuits.

References

[1] International Telecommunications Union Recommendation ITU.372-8 (2005).

[2] Poisel, R. A., *Modern Communications Jamming Principles and Techniques*, Norwood, MA: Artech House, 2004, p. 15.

[3] Skomal, E. N., *Man-Made Radio Noise*, New York: Van Nostrand, 1978.

[4] Riley, N. G., and K. Docherty, "Modeling and Measurement of Man-Made Radio Noise in the VHF–UHF Bands," *IEEE Antennas and Propagation Symposium*, April 4–7 1995, pp. 313–316.

[5] Parsons, D., *The Mobile Radio Propagation Channel*, New York: Wiley, 1992, Ch. 9.

[6] Gagliardi, R. M., *Introduction to Communications Engineering*, 2nd Ed., New York: Wiley, 1988, p. 151.

[7] Lathi, B. P., *Communication Systems*, New York: Wiley, 1968, Chapter 6.

[8] Mehrotra, A., "Noise in Radio Frequency Circuits: Analysis and Design Implications," *Proceedings 2001 International Symposium on Quality Electronic Design*, March 26–28, 2001, pp. 469–476.

[9] Wyatt, J. L., and G. J. Coram, "Nonlinear Device Noise Models: Satisfying the Thermodynamic Requirements," *IEEE Transactions on Electron Devices*, Vol. 46, No. 1, January 1999, pp. 184–193.

[10] Roychowdhury, J., and A. Demir, "Estimating Noise in RF Systems, *IEEE/ACM International Conference on CAD, Digest of Technical Papers*, 1998, pp. 199–202.

[11] Hull, C. D., and R. G. Meyer, "A Systematic Approach to the Analysis of Noise in Mixers," *IEEE Transactions on Circuits and Systems–I: Fundamental Theory and Applications*, Vol. 40, No. 12, December 1993, pp. 909–919.

[12] Rizzoli, V., F. Mastri, and D. Masotti, "General Noise Analysis of Nonlinear Microwave Circuits by the Piecewise Harmonic-Balance Technique," *Transactions on Microwave Theory and Techniques*, Vol. 42, No. 5, May 1994, pp. 807–814.

[13] Niu, G., "Noise in SiGe HBT RF Technology: Physics, Modeling, and Circuit Implications," *Proceedings of the IEEE*, Vol. 93, No. 9, September 2005.

[14] Van Der Ziel, A., "Noise in Solid-State Devices and Lasers," *Proceedings of the IEEE*, Vol. 58, No. 8, August 1970, pp. 1178–1206.

[15] Van Der Ziel, A., "Unified Presentation of 1/f Noise in Electronic Devices: Fundamental 1/f Noise Sources," *Proceedings of the IEEE*, Vol. 76, No. 3, March 1988, pp. 233–258.

[16] Hooge, F. N., "1/f Noise Sources," *IEEE Transactions on Electronic Devices*, Vol. 41, No. 11, November 1994, pp. 1923–1935.

[17] Keshner, M. S., "1/f Noise," *Proceedings of the IEEE*, Vol. 70, No. 3, March 1982, pp. 212–218.

[18] Leeson, D. B., "Simple Model of Feedback Oscillator Noise Spectrum," *Proceedings of the IEEE*, February 1966, pp. 329–330.

[19] Reimbold, G., "Modified 1/f Trapping Noise Theory and Experiments in MOS Transistors Biased from Weak to Strong Inversion Influence of Interface States," *IEEE Transactions on Electronic Devices*, Vol. ED-31, No. 9, September 1984, pp. 1190–1198.

[20] Oksendal, B., *Stochastic Differential Equations*, Oslo: Springer, 2003.

[21] Kleinpenning, T. G. M., and A. H. de Kuijper, "Relation between Variance and Sample Duration of *1/f* Noise Signals," *Journal of Applied Physics*, Vol. 6, p. 43, 1988.

[22] Lee, T. H., and A. Hajimiri, "Oscillator Phase Noise: A Tutorial," *IEEE Journal of Solid State Circuits*, Vol. 35, No.3, March 2000, pp. 326–336.

[23] Demir, A., and J. Roychowdhury, "Phase Noise in Oscillators: A Unifying Theory and Numerical Methods for Characterization," *IEEE Transactions on Circuits and Systems–I: Fundamental Theory and Applications*, Vol. 47, No. 5, May 2000, pp. 655–674.

[24] Mini-Circuits, Inc., "Characterizing Phase Noise," *RF Design*, January 2003, pp. 58–59.

[25] Smith, P., "Little Known Characteristics of Phase Noise," *RF Design*, March 2004, pp. 46–52.

[26] Goldberg, G.–G., "Oscillator Phase Noise Revisited–A Heuristic Review," *RF Design*, January 2002, pp. 52–64.

[27] Candy, J. C., and G. C. Temes. "Oversampling Methods for A/D D/A Conversion, Oversampling Delta-Sigma Converters," *IEEE Press*, New Jersey, 1992, pp. 2–3.

[28] Goldberg, B. –G., "The Effects of Clock Jitter on Data Conversion Devised," *RF Design*, August 2002, pp. 26–32.

[29] Parker, W. H., "Electromagnetic Interference: A Tutorial," *Proceedings IEEE Aerospace Applications Conference* 1996, Vol. 3, February 3–10 1996, pp. 177–183.

[30] Terman, F. E., *Radio Engineers Handbook*, New York: McGraw-Hill, 1943, pp. 30–37.

[31] Rheinfelder, W. A., *Design of Low-Noise Transistor Input Circuits*, New York: Hayden, 1964.

[32] Israelsohn, J., "Noise 101," *EDN*, January 8, 2004, pp.41–47.

[33] Israelsohn, J., "Noise 102," *EDN*, March 18, 2004, pp. 46–54.

[34] Martin, S., V. D. Archer III, D. M. Boulin, M. R. Frei, K. K. Ng, and R.–H Yan, "Device Noise in Silicon RF Technologies," *Bell Laboratories Technical Journal*, September 1997, pp. 30–45.

[35] Leach, W. M., "Fundamentals of Low-Noise Analog Circuit Design," *Proceedings of the IEEE*, Vol. 82, No. 10, October 1994, pp. 1515–1538.

[36] Nguyen, T.–K., C.–H. Kim, G.–J. Ihm, M.–S. Yang, and S.–G. Lee, "CMOS Low-Noise Amplifier Design Optimization Techniques," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 52, No. 5, May 2004, pp. 1433–1442.

[37] Roa, E., J. N. Soares, and W. Van Noije, "A Methodology for CMOS Low Noise Amplifier Design," *Proceedings of the 16th IEEE Symposium on Integrated Circuits and Systems Design*, 2003.

[38] Motchenbacher, C. D., and J. A. Connelley, *Low-Noise Electronic System Design*, New York: Wiley, 1993.

[39] Friis, H. T., "Noise Figures of Radio Receivers," *Proceedings of the IRE*, Vol. 32, July 1944, p. 418.

Chapter 6

Radio Communication Technologies

6.1 Introduction

An understanding of communication technologies is important to understand how to design EW systems. These technologies facilitate the exchange of information between two entities, and therefore are the means by which information is passed. Information not shared is useful, but only to the entity that has possession of it. It is far more useful and much more can be done with information if it is shared. Thus, information transport is a vital part of any information-based society, including the military.

There are many good texts available that discuss communication technologies and systems at length. This chapter is included to provide just an introduction to the subject at a depth necessary to understand what is presented here. It is not intended as a thorough review of the subject. For more in-depth reading, the reader is referred to [1–6].The generic information transport system model used herein is shown in Figure 6.1. There is a source that generates information that is to be transferred to one or more sinks. There are several ways to make a distinction between the techniques for accomplishing this information transfer, but the fundamental distinction made here is whether the transfer medium is a wire and/or fiber or wireless. Note that information storage, on a floppy disk, for example, is another form of information transport system where the information is to be transferred primarily over the time dimension.

There is usually channel noise associated with any communication system, regardless of the communication medium. Here it is assumed that all noise is additive. Some of those that are more important to communication EW system design were discussed in Chapter 5. Noise is any signal that is other than the intended signal. The communication system in Figure 6.1 will be discussed in this chapter. The remainder of the book will discuss in detail the jammer, the intercept system, and their performance.

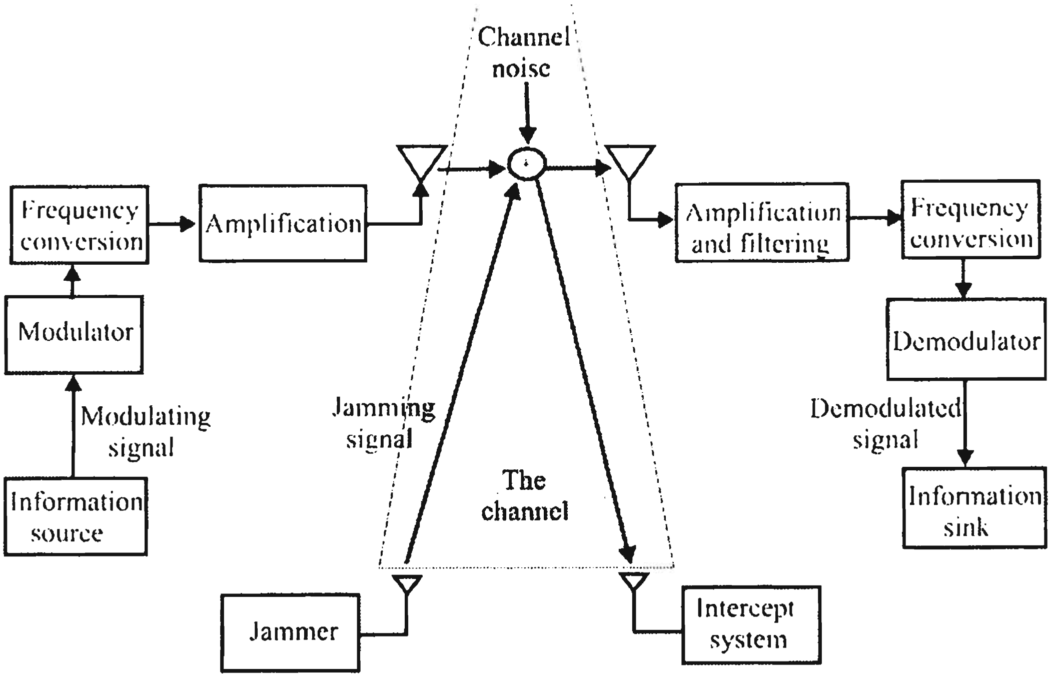


Figure 6.1 Generic communication system model with a jammer and intercept system in the channel.

Communication systems can be categorized as either broadcast or netted. These appellations, however, are not meant to imply a size. Either can be small or very large. Broadcast communications are usually generated at a single place and sent to many receivers. Examples of this are commercial radio and TV broadcasts. In netted communications, on the other hand, there are several possible sources of information that is to be sent to other locations. Perhaps every communication node on the network can generate such information. The Internet is an example of a large netted network. Military, mobile tactical command and control is normally implemented with networks of VHF and UHF communication radios. Figure 6.2 illustrates the fundamental differences in broadcast versus netted communications.

Either of these communication models can be encountered in military situations. Communication EW systems can be expected to operate against either. An example of a broadcast EW application is when the former Soviet Union jammers denied radio-free Europe to most of Eastern Europe. An example of EW against tactical networks is classical VHF jamming.

Frequently in a tactical mobile network there is one node in the network that is referred to as the net control station (NCS). Normally it would be the node associated with the person in charge of the unit using the radios. There is often more communication traffic from the NCS than the other nodes. In some cases that can be used to identify the NCS.

The remainder of this chapter is structured as follows. Section 6.2 introduces the modulation techniques used in modern communication systems. That is followed by a

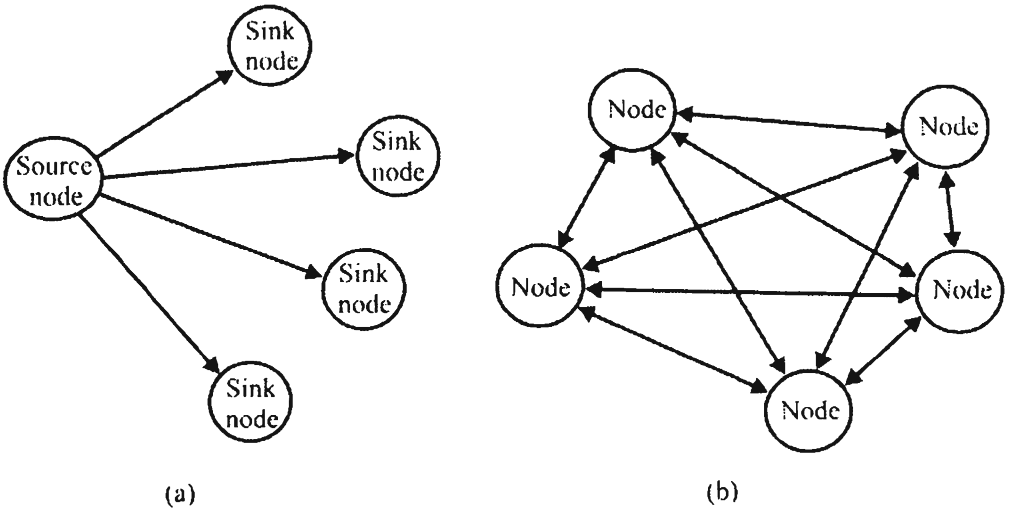


Figure 6.2 Two fundamental types of communications: (a) broadcast and (b) networked.

discussion of methods to access the RF spectrum so that multiple users can share it. The introduction of duplexing techniques follows that. Duplexing is the methodology for two users to communicate back and forth. Most modern digital communication systems employ coding for a couple of reasons; coding is discussed next. That is followed by a discussion of modems and their protocols. The topic of facsimile communications comes after that. The chapter concludes with an introduction to encryption technologies.

6.2 Modulation

Modifying a carrier signal with the information to be transported is called modulation. The reason for putting the information-bearing signal on a separate carrier is to move the information signal to a region of the frequency spectrum where transmission is more economical. (As the frequency gets higher, the antenna gets smaller, thereby requiring less real estate. There are other economic reasons as well.)

There are three fundamental ways that a radio signal can be modified to carry information: its amplitude, its frequency, and its instantaneous phase angle. See Figure 6.3. These are referred to as *amplitude modulation* (AM), *frequency modulation* (FM), and *phase modulation* (PM)—the last two are referred to as angle modulations.

The signal that is modified by modulation is called the carrier signal, while the information-bearing signal, which does the modification, is called the modulating signal denoted by $m(t)$. The modulated carrier signal is given by

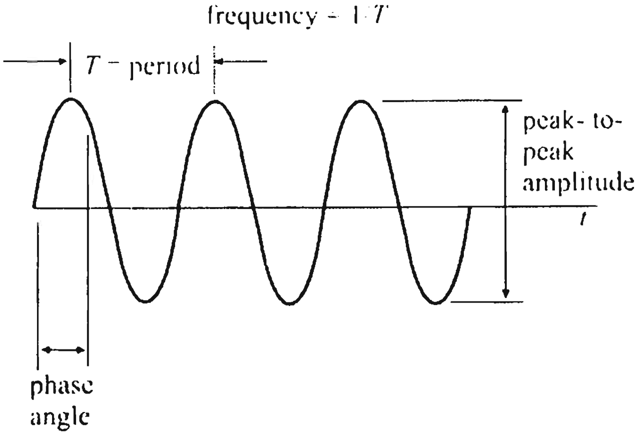


Figure 6.3 Fundamental parameters of an EM sine wave.

$$s(t) = A(t)\cos[\varphi(t)] = A(t)\cos[2\pi f_c t + \Psi_c + \phi(t)] \qquad (6.1)$$

where $A(t)$ is the amplitude of the signal, f_c is the carrier frequency, Ψ_c is the initial phase of the carrier signal, and $\phi(t)$ is an additional phase term described shortly. AM modifies $A(t)$, while phase and frequency modulation modifies $\varphi(t)$ by modifying $\phi(t)$.

The modulation used to convey commercial radio signals in the United States on the carrier frequencies between 550 kHz–1.5 MHz, well-known to all, is AM. In that case the amplitude of the carrier is modified. For the FM radio case, where the commercial broadcast frequencies are 88–108 MHz, the modulation technique is referred to as FM. In this case it is the frequency that is changed by the information-bearing signal. The third parameter that can be changed is the phase of the carrier. The phase is measured relative to a reference at the receiver, so it is relative phase that is important. In addition, there are limits to how much the phase can be changed. Any phase change larger than 2π radians would lead to ambiguities, since $\theta = n2\pi + \theta$ for any integer n.

Work on modulation techniques has long been focused on achieving spectrum efficiency, that is, putting more information through a given amount of bandwidth. One scheme is to reduce the redundancy in human speech and writing. For example, the letter u (almost) always follows the letter q, so why transmit the u? The analog AM and FM communication technologies generally attain much less than the equivalent of 1 bit per second per hertz of bandwidth. Modulating carriers with digital information using these analog schemes are no better. On the other hand, with PM, 4 bps/Hz of RF bandwidth and higher have been achieved, thus explaining the attractiveness of this form of modulation. This, however, does not come freely. Synchronizing sequences and circuitry are required between the transmitter and receiver in order to establish and maintain communications. Furthermore, normal characterization of the communication channel is required in order to prevent one symbol from interfering with the next. Appropriate processing of the communication signal is subsequently required.

Demodulation is the inverse process—removing the modulating signal from the carrier. Demodulation is different from detection, the latter of which is used in two ways. In communication terminology, detection refers to the determination of which of several symbols was transmitted in digital signaling. For communication EW system design, detection most of the time refers to the determination of whether there is a signal present at a channel or not.

There are two fundamental forms of demodulation. The first is noncoherent demodulation. In that case, no carrier phasing information is used in the demodulation process. For coherent demodulation, the carrier phase information is used. The latter generally yields better performance. However, it is often accompanied by more complex hardware implementations. Performance in this case refers to signal detection.

For all of the PM schemes, the receiver needs to know where the reference carrier phase is located. Thus synchronization of the receiver with the incoming signal and coherent detection are required. Recovery of the symbol clock is required in all of the demodulation schemes. Coherent detection is possible with other modulation schemes as well, and when it can be done, it typically provides for 3 dB or more improvement in sensitivity performance.

6.2.1 Analog Modulations

6.2.1.1 Amplitude Modulation

The process of amplitude-modulating a carrier signal is illustrated in Figure 6.4. If the modulating signal is a tone signal (single-frequency, zero-bandwidth), then the spectrum of the modulated carrier is as shown in Figure 6.5. For real baseband waveforms that have some finite nonzero bandwidth, the spectrum appears as shown in Figure 6.6.

Figures 6.5 and 6.6 illustrate the case of *double sideband* (DSB) AM, where substantial portions of the carrier are transmitted with the information part of the signal. If the carrier is suppressed (either completely or partially), it is called DSB *suppressed carrier* (SC). If the upper sideband of the signal is suppressed as well, then the signal is one of the *single sideband* (SSB) forms called *lower sideband* (LSB). Likewise if the lower sideband is suppressed, this form of SSB is called *upper sideband* (USB). One of the sidebands can be suppressed because for AM each of the

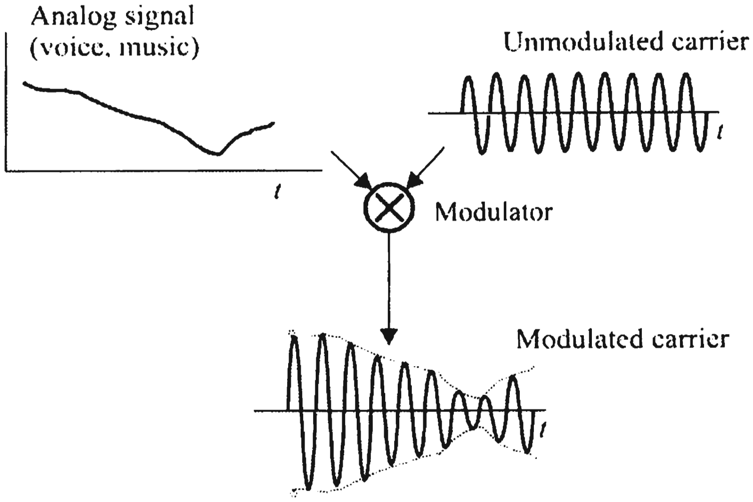


Figure 6.4 AM varies the amplitude of the carrier waveform.

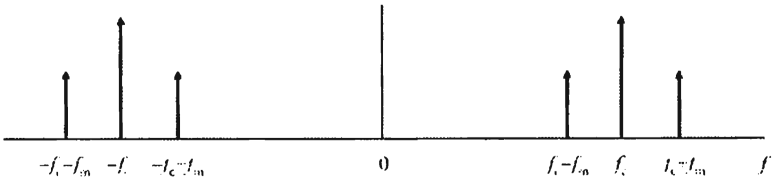


Figure 6.5 Amplitude spectrum of a signal that is AM-modulated with a tone.

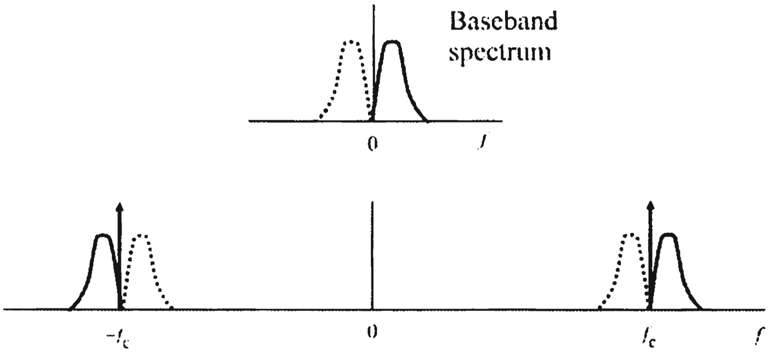


Figure 6.6 Spectrum of AM waveform when the modulating waveform has nonzero bandwidth.

sidebands contains the same information. Suppression of a sideband reduces the required bandwidth for transmission. As seen in Figure 6.5, the bandwidth of the modulated carrier is either 1 × baseband bandwidth, or 2 × baseband bandwidth. The former is when the modulation is AM SSB and the latter is for AM DSB.

Atmospheric noise affects AM more than other forms of modulation. This is particularly true in the lower-frequency ranges where lightning causes substantial interference to AM signals.

6.2.1.2 Angle Modulation

The instantaneous phase of the carrier is given by

$$\varphi(t) = 2\pi f_c t + \Psi_c + \phi(t) \tag{6.2}$$

while the instantaneous frequency is

$$f_i = \frac{1}{2\pi}\frac{d\varphi(t)}{dt} = f_c + \frac{1}{2\pi}\frac{d\phi(t)}{dt} \tag{6.3}$$

for PM

$$\phi(t) = k_p m(t) \tag{6.4}$$

where k_p is the PM constant given in radians/volt, or perhaps radians/ampere. For FM

$$\phi(t) = k_m \int_{-\infty}^{t} m(\tau)d\tau \tag{6.5}$$

where k_m is the FM constant given in radians/second/volt or radians/second/ampere.

The modulated carrier waveforms, then, for PM and FM are (ignoring Ψ_c)

$$\text{PM:} \quad s(t) = A\cos[2\pi f_c t + k_p m(t)] \tag{6.6}$$

$$\text{FM:} \quad s(t) = A\cos[2\pi f_c t + k_f \int_{-\infty}^{t} m(\tau)d\tau] \tag{6.7}$$

Frequency-modulating a signal has the effect shown in Figure 6.7. The instantaneous frequency is changed. The effects in the frequency domain are more complex than in

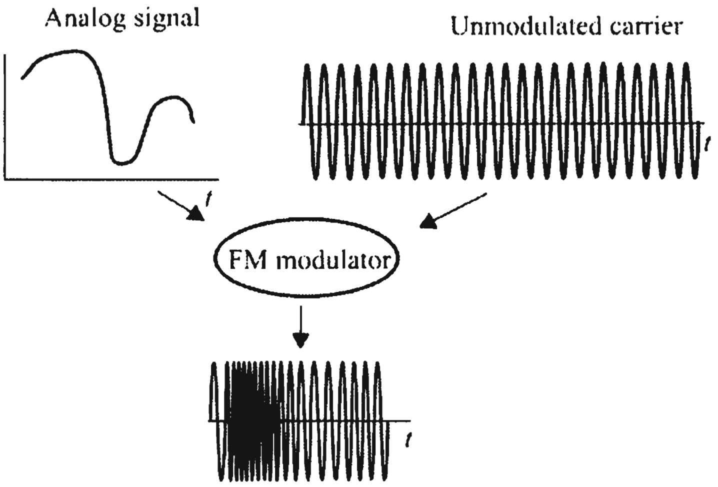


Figure 6.7 FM of a carrier varies its instantaneous frequency around the carrier frequency.

the AM case above because FM is a nonlinear modulation technique. If the modulating signal is a tone, then the spectrum of the modulated carrier is like that shown in Figure 6.8. The spacing of the frequency lines of the modulated signal is equal to f_m. If the modulating signal is not a tone, then the spectrum is very complicated and in most cases has yet to yield to closed form mathematical analysis.

Let $m(t) = M\cos(2\pi f_m t)$ where M is a constant amplitude so that the modulating signal is also a sinusoidal signal such that $f_m << f_c$. Then for FM

$$s(t) = A\cos[2\pi f_c t + \beta\sin(2\pi f_m t)] \tag{6.8}$$

$\beta = k_m M/2\pi f_m$ is the modulation index. β is the ratio of the maximum frequency

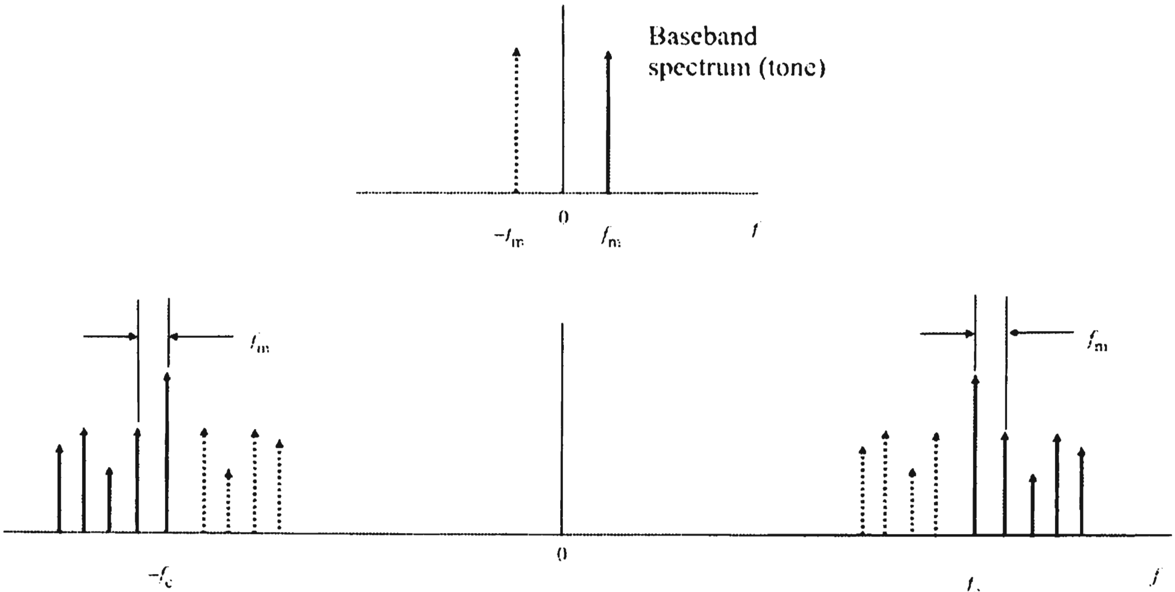


Figure 6.8 Spectrum for FM where the modulating signal is a tone.

deviation of the carrier caused by the modulating signal, to the maximum frequency of the modulating signal.

The cosine term can be expanded as

$$\begin{aligned}\cos[2\pi f_c t+\beta\sin(2\pi f_m t)] = J_0(\beta)\cos(2\pi f_c t)&\\ +\sum_{n=1}^{\infty}(-1)^n J_n(\beta)[\cos(2\pi f_c - n2\pi f_m)t]&\\ +(-1)^n\cos(2\pi f_c + n2\pi f_m)t]&\end{aligned} \tag{6.9}$$

where the J_i are Bessel functions of the first kind. For small β,

$$J_0(\beta)\approx 1-\left(\frac{\beta}{2}\right)^2 \tag{6.10}$$

$$J_n(\beta)\approx \frac{1}{n!}\left(\frac{\beta}{2}\right)^n,\ n\neq 0 \tag{6.11}$$

Thus,

$$s(t)\approx A\cos(2\pi f_c t)-\frac{A\beta}{2}\cos\left[2\pi(f_c - f_m t)\right]+\frac{A\beta}{2}\cos\left[2\pi(f_c + f_m)t\right] \tag{6.12}$$

which simplifies to

$$s(t)\approx A\cos(2\pi f_c t)-A\beta\sin(2\pi f_m t)\sin(2\pi f_c t) \tag{6.13}$$

When $\beta < 1$, it is called narrowband FM, and the bandwidth of the signal is about 2 × f_m. The spectrum of this signal is shown in Figure 6.9 (taking into consideration the negative frequencies as well).

FM is a nonlinear modulation, as can be ascertained by (6.7)–(6.11). For AM, which is a linear modulation, determination of the bandwidth of the signal was straightforward—it was either 1 × baseband bandwidth or 2 × baseband bandwidth. For FM it is not so straightforward. Carson developed an approximation for the bandwidth of FM signals, which is known as the Carson rule. It states that the bandwidth is given by

$$B_{IF}\approx 2(\Delta F+f_m) \tag{6.14}$$

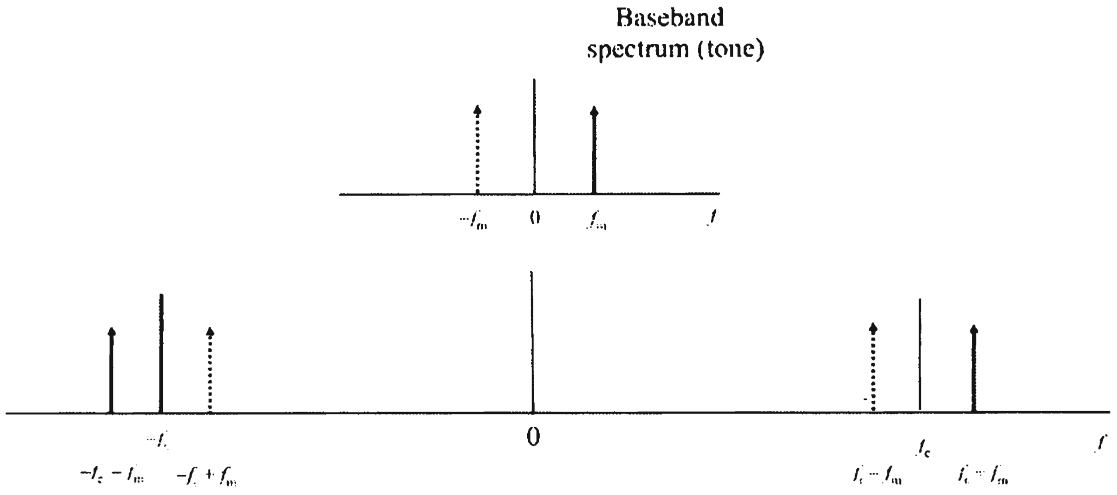


Figure 6.9 Frequency spectrum of narrowband FM where the modulating signal is a tone.

where β = modulation index = $\Delta F/f_m >> 1$; B_{IF} = IF bandwidth; ΔF = peak carrier frequency deviation; and f_m = peak-modulating frequency deviation.

An extension to Carson's rule was subsequently developed for smaller modulation indices and is given by

$$B_{IF} = 2(\Delta F + 2f_m) \tag{6.15}$$

when $2 < \beta < 10$.

An effective and simple FM modulator is made with a voltage-controlled oscillator (VCO). Such a configuration is shown in Figure 6.10. The characteristics of the VCO are shown at the top of Figure 6.10. The input voltage (slowly changing

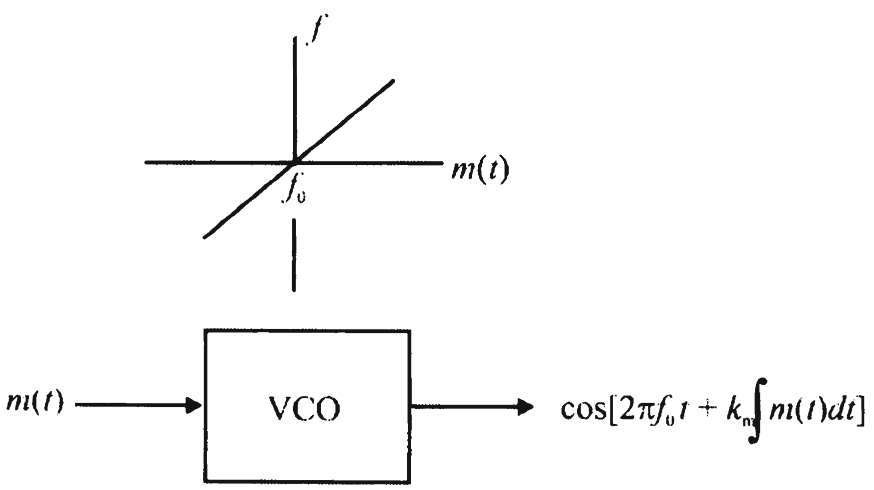


Figure 6.10 An FM modulator constructed with a VCO. The VCO transfer characteristic is shown at the top.

relative to the oscillator's frequency) controls the frequency of oscillation. Within its design range, this is a linear relationship.

For most common FM cellular systems $1 < \beta < 3$, so Carson's rule does not apply very well (it is not a very good approximation). For the original North American analog cell phone standard, the Advanced mobile phone system (AMPS), $\beta \approx 2.7$, producing detected SNRs, denoted SNR_d, of about 40 dB from a predetected carrier-to-noise ratio, denoted by CNR_IF, of 18 dB.

The derivation of the spectrum for PM is similar to the above for FM and many of the same results apply. Carson's rule for PM is

$$B_\text{IF} = 2(\beta+1)f_\text{m} \qquad \beta > 10 \tag{6.16}$$

$$B_\text{IF} \approx 2(\beta+2)f_\text{m} \qquad 2 < \beta < 10 \tag{6.17}$$

where

- β = PM modulation index = $\Delta f / f_\text{m}$
- f_m = highest modulating frequency.

Two baseband signals can be sent at the same time using the same carrier using orthogonal signaling, also called quadrature modulation. Two functions are *orthogonal* if their inner product is zero. In the case of importance here, two periodic functions defined over $(0, nT)$ with period T and integer n given by $f(t)$ and $g(t)$ are orthogonal if and only if

$$\int_0^{nT} f(t)g(t)dt \begin{cases} \neq 0 & \text{if} \quad f(t) = g(t) \\ = 0 & \text{if} \quad f(t) \neq g(t) \end{cases} \tag{6.18}$$

Specific examples of important orthogonal waveforms in communications are $\sin 2\pi ft$ and $\cos 2\pi f t$.

Let $r(t)$ represent the received RF waveform. Furthermore, let $r(t) = A\sin(2\pi f t) + B\cos(2\pi f t)$. Here A and B are two signals to be transmitted that may or may not be related to each other. A and B are treated as constants normally associated with digital communications, either $A, B \in \{0, 1\}$ or, more commonly $A, B \in \{1, -1\}$. Now suppose at the receiver, $r(t)$ is divided into two paths as shown in Figure 6.11. In one path is a multiplier where $r(t)$ is multiplied by $\sin(2\pi f t)$, while in the other path it is multiplied by $\cos(2\pi ft)$. These multipliers are followed by integrators where the period of integration is a multiple of the period of the carrier period $T = 1/f$. Then the output of the top integrator will be

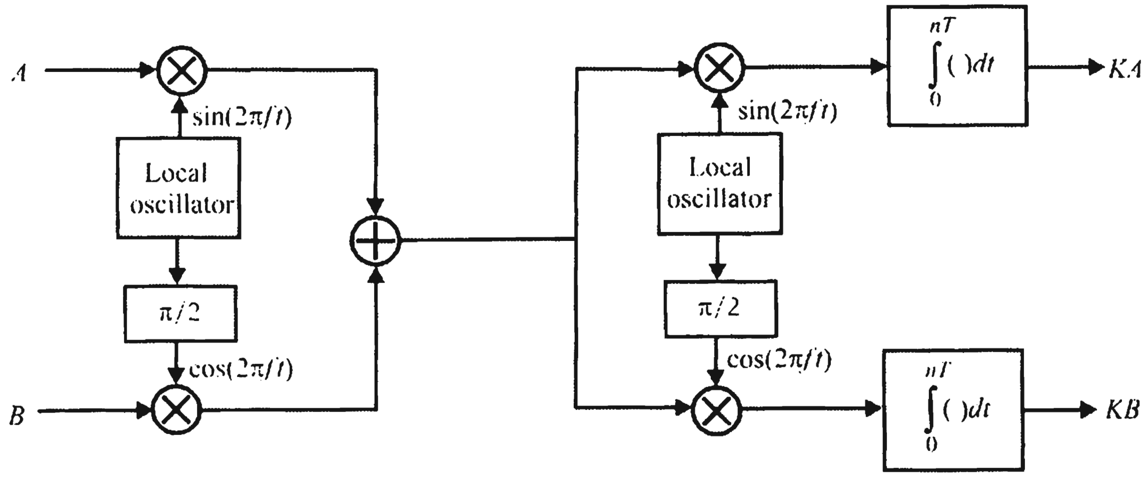


Figure 6.11 Orthogonal modulator and demodulator allow two distinct signals to share the same carrier.

$$\begin{aligned}
\int_0^{nT} r(t)\sin(2\pi ft)dt &= \int_0^{nT} A\sin^2(2\pi ft)dt + \int_0^{nT} B\cos(2\pi ft)\sin(2\pi ft)dt \\
&= \frac{A}{2}\int_0^{nT}\left[1-\cos(2\pi ft)dt\right] + \frac{B}{2}\int_0^{nT}[\sin(4\pi ft)dt - \sin(0)]dt \\
&= \frac{A}{2}\int_0^{nT} dt - \frac{A}{2}\int_0^{nT}\cos(2\pi ft)dt \\
&\qquad + \frac{B}{2}\int_0^{nT}\sin(4\pi ft)dt - \frac{B}{2}\int_0^{nT} 0dt \\
&= \frac{AnT}{2} = KA
\end{aligned} \tag{6.19}$$

All of the integrals but the first one are zero because any zero-mean periodic function integrated for an integer number of periods is zero, and, of course, the last integral is zero anyway. The second integral in the first line of (6.15) is equal to zero by inspection due to the orthogonality property discussed above. It can similarly be shown that the output of the bottom integrator is equal to *BnT*/2. Therefore, the outputs of the integrators are a constant multiple of the baseband waveforms.

This configuration does impose some requirements on the communication system. For example, the transmitter and receiver local oscillators must be synchronized so they are in phase coherence with one another. If they are not in phase coherence and/or if the integrals are not exactly an integer multiple of periods, the outputs are not exactly equal to a constant times the input, but will be somewhat different. That is, some noise is introduced into the demodulation process. When used for digital signaling, this is not too critical since it only becomes a problem when the transmitted symbol was an A, for example, and the other possible symbol is $-A$. If the demodulated symbol is closer to $-A$ than A, a symbol error will occur. The lower the SNR, the more likely this possibility will be.

If $A = a(t)$ and $B = b(t)$ are not constant, but functions of time, then this analysis and configuration still works. However, if $a(t)$ and $b(t)$ are the information-bearing signals, then it is necessary to take their derivatives at the transmitter before being submitted to the modulation process. That way the output of the integrators at the receiver will be functions of the baseband signals, not their integral as they would be otherwise.

6.2.2 Digital Modulations

Transmission of signals using digital signaling techniques has spread considerably in the past several decades [7]. The reasons are simple: generally signals that are modulated by digital methods are more spectrally efficient that those that are analog-modulated. In addition coding can be applied to digital signals in order to compensate for imperfections in the channel. The carrier for digital signaling can be either amplitude-, frequency-, or phase-modulated, or combinations.

The signals to be transmitted may or may not be natively digital. Computer-to-computer information exchange is digital from birth. Analog telephone conversations are not, however, and the analog voice signals must first be converted to digital form in order to take advantage of the benefits digital signaling provides.

Wireline communication, such as that provided by the public-switched telephone network (PSTN), is a fairly benign environment. Induced noise is low and interfering signals can be controlled. Communication by radio is a different story. It is a much harsher environment in that external noise is higher, interfering signals cannot normally be controlled, and knowing the propagation parameters is difficult at best.

Two signals $s_1(t)$ and $s_2(t)$ are *antipodal* if $s_2(t) = -s_1(t)$. Thus $\{-1, +1\}$ form an antipodal set as do the two signals shown in Figure 6.12. Two signals are *coherent* with one another if they maintain a known phase relationship with each other and two signals are *noncoherent* with one another if there is no consistent or known phase relationship between the two.

6.2.2.1 Pulse Code Modulation

In order to transmit analog information, such as human speech and television images, over digital channels, the signals must first be converted into digital form. This is normally done with analog-to-digital (A/D) converters. A/D converters change an analog signal into digital representations of those signals. These digital

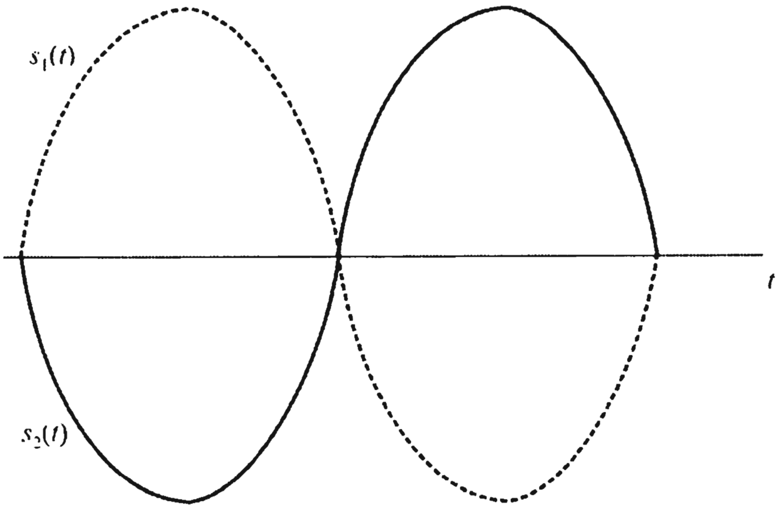


Figure 6.12 Antipodal signal set.

representations are referred to as pulse codes in the domain of digital communications. The output of an A/D converter is simply the (discrete) binary approximation of the amplitude of the voltage or current at the input to the A/D at each sample time. When these digital words are modulated onto a carrier, the result is called *pulse code modulation* (PCM). This modulation can take many forms, some of which are discussed next.

Differential PCM

PCM assigns a number in digitized form to the amplitude of an analog signal. If the signal is not changing too rapidly from one sample to the next, there will be considerable redundancy from one sample to the next (for example, there is not too much difference between 1 and 1.01V). An attempt to remove this redundancy was developed, called *differential PCM* (DPCM). In DPCM, instead of encoding every sample with the full number of bits of the A/D converter, there are fewer than the total number of bits used; frequently only 1-bit difference from one sample to the next is used. That bit is +1 if the next sample is larger than the current sample, and it is –1 if the next sample is smaller.

Adaptive Differential PCM

A variant of DPCM is *adaptive differential PCM* (ADPCM). It is similar to DPCM except that if the next sample is substantially larger than the current sample, there are provisions for using more than +1 or –1 to encode the change. Figure 6.13 shows how ADPCM works. Each digitized sample of an analog sample is compared with the last sample. While for DPCM, if the sample is larger, +1 (bit) is used to represent the

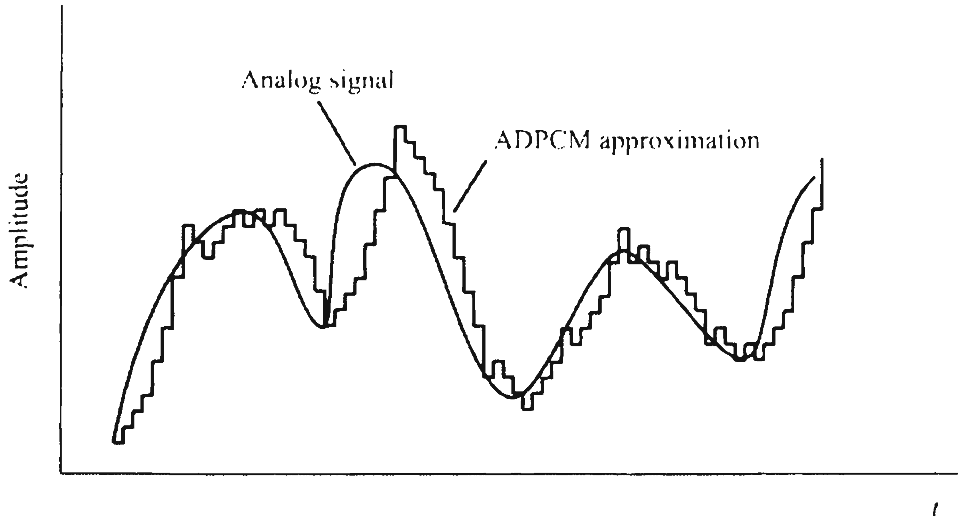


Figure 6.13 ADPCM approximating the analog signal.

change and if it is smaller, a −1 (bit) is used to represent the change, as seen in Figure 6.13 more than 1 bit is used in ADPCM. There are only +1 bits and −1 bits, no zero bits, so if there is no change in the signal, +1 and −1 will alternately be coded. This can be seen in Figure 6.13 where the analog signal is reversing directions, and therefore, the changes are small enough to be less than the difference in one sample. In ADCPM, where, after a certain number of attempts (three in this example), the digital samples fail to catch up with the analog waveform, the size of the step is increased (by one step here) until the analog waveform is matched or exceeded. At that point, the polarity of the step is reversed and the amplitude reduced back to one step size.

Since changes larger than 1 bit are used, more than 1 bit is necessary to encode the difference. For example, up to four levels can be encoded with two binary bits. ADPCM allows the digitized waveform to catch up faster.

6.2.2.2 Power Efficiency

Power efficiency is determined by the required energy per bit to noise level, as specified, for example, by the SNR in watt-hours per bit per hertz of bandwidth, or E_b/N_0, in order to produce a specified BER. Over a ground-to-ground tactical army communication channel, noise and interference dictates that a BER of 10^{-2} is typical, whereas over a satellite link the BER is typically 10^{-5} or better. The BER is a system design parameter and many techniques can be used to achieve low BER even over noisy, normally high BER channels. Usually an additive white gaussian noise (AWGN) channel is assumed, especially for satellite links. In this type of link, the noise is assumed to behave according to a Gaussian pdf and such noise is added to any signal being sent through it.

6.2.2.3 Bandwidth Efficiency

Bandwidth efficiency is defined as the number of bits per second that can be transmitted over 1 Hz of bandwidth. The higher the bandwidth efficiency, the more information that can be transmitted over that channel. The amount of frequency spectrum is fixed—there is only a certain amount of it. In order to increase the amount of communication once there are signals traveling over every channel in a given area of the spectrum, it is necessary to increase the bandwidth efficiency. Modulation techniques must be bandwidth-efficient to allow the maximum number of users possible to use the propagation medium.

6.2.2.4 Other Considerations for Digital Communications

As discussed in Section 3.5, multipath is the simultaneous reception of two or more copies of a signal. It is caused by reflections of the signal from potentially one or more objects. Some forms of modulations are more tolerant to multipath effects than others, so in cases where toleration of multipath is required, those methods may be appropriate.

Another form of interference is when other signals are impinging on the receiver at the same time as the intended signal. This is called cochannel interference and some forms of signal processing are more tolerant of it than others.

Economics always is a consideration in all but the most exceptional cases. Therefore, initial investment in the communication system is usually a concern. In most cases, however, cost of ownership involves much more than the initial investment. Operation and support (O&S) costs typically run much higher than the initial investment. Minimizing those can often be of great importance. An example of this is for cellular systems. If there were no rechargeable batteries to reduce the cost of powering the handsets, it is unlikely that modern cellular systems would be economically viable.

When the RF spectrum is tightly packed, as all the lower RF communication bands are today, out-of-band spillover of energy is a consideration. This is accidentally putting the energy from one channel into adjacent channels. Of course, some spillover is inevitable in any practical implementation. Some schemes are more tolerant than others, however.

The narrower the main lobe width in the frequency spectrum, the more spectrally efficient the modulation scheme. This is not the whole story, however, as sidelobes of energy can show up in adjacent channels, as well as channels far away from the intended channel. While each of the modulation schemes exhibits its own unique characteristics, it is certainly a goal of them all to keep the energy of the signal all within the desired channel. Not only do the sidelobes interfere directly with the signals in the other channels, but also the energy spilled over interacts with the signals

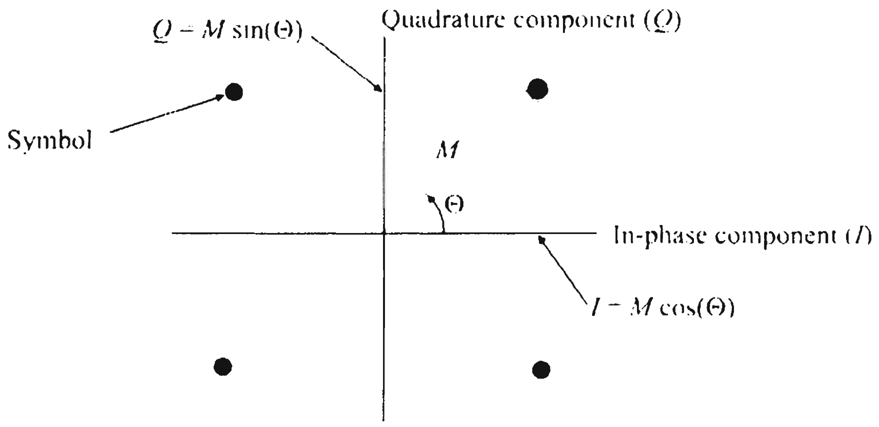


Figure 6.14 AM/ PM constellation.

in the sidelobes. When the system is nonlinear, this interaction can generate intolerable intermodulation interference. Spillover energy is wasteful, as is the energy in the intermodulation products, exacerbating the power inefficiency.

6.2.2.5 Constellations of Amplitude and Phase Modulations

Most of the digital modulation schemes described herein can be visualized with the aid of the constellation. An example of a constellation is shown in Figure 6.14. Both the phase and amplitude, in general, can and are modulated (changed) with these modulations. By convention, the component of the vector that is in phase with the carrier (before modulation) is called the in-phase component, while the component on the other axis is called the quadrature component. Each combination of amplitude and phase (in phasor notation) or in-phase and quadrature in x, y dimensions is called a symbol.

6.2.2.6 Amplitude Shift Key

When the amplitude of the carrier signal is changed between two (or more) levels, frequently either A and 0, then the carrier is *amplitude shift key* (ASK)-modulated. The resultant signal is as shown in Figure 6.15 along with a simple notional implementation. When one of the levels is zero, this type of modulation is also referred to as *on-off keying* (OOK). This is the form of modulation originally invented by Morse for use with the telegraph.

OOK is the most spectrally polluting form of amplitude modulation for data transmission. Simply altering the bias or controlling the supply to the oscillator will achieve this. The spectrum gets polluted because turning the carrier on and off with a square wave modulating signal makes the spectrum of the modulated signal extend many times beyond the data rate being conveyed.

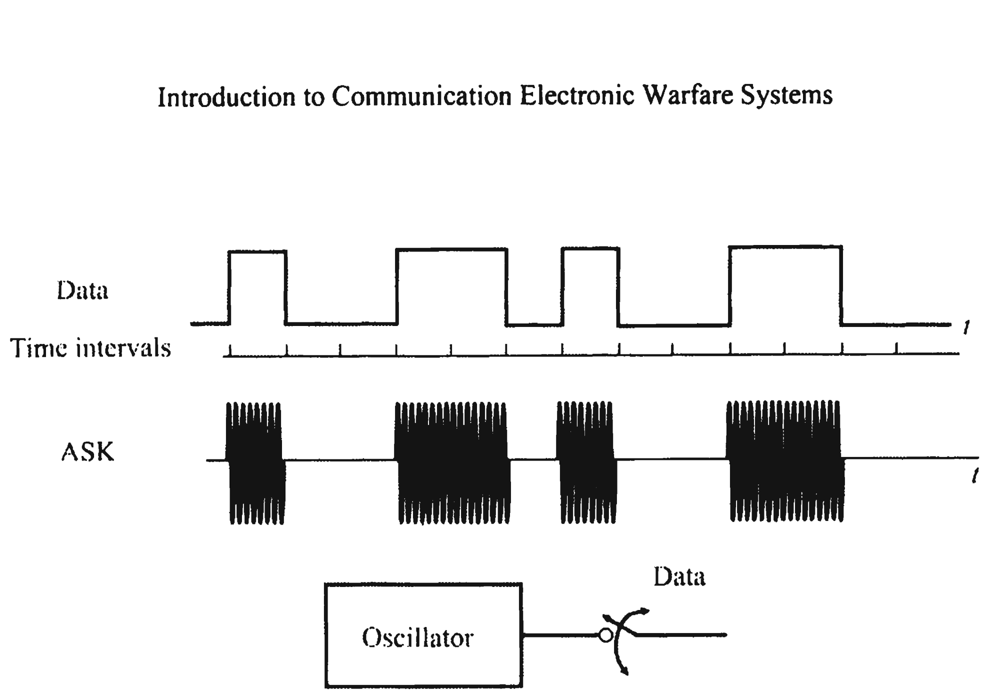


Figure 6.15 ASK modulation.

Controlling the turn on and turn off rate of the carrier so that the amplitude grows and decays smoothly significantly reduces spectral pollution without degrading the performance of the system. While it is possible to design a circuit that will achieve this for the carrier source itself, it is often simpler and cheaper to use a mixer to impose a preshaped data signal onto the fixed amplitude carrier.

6.2.2.7 Frequency Shift Key

In *frequency shift key* (FSK) modulation the carrier signal is changed between two (or more) frequencies according to the digital signal as shown in Figure 6.16. It is notionally implemented by switching rapidly between the two frequencies as shown. Alternately, the switching could be between two baseband frequencies and the subsequent signal, consisting of f_{m1} and f_{m2} used to frequency-modulate the carrier. Note that, in general, when the data changes, the phase of the carrier is discontinuous, unless mechanisms are put into place to control it.

Vector Modulator

A *vector modulator*, also called a *quadrature modulator*, implements a complex mathematical sum, multiplying two input signals, $\pm\cos \omega_m t$ and $\sin \omega_m t$, where ω_m is the desired frequency shift to be imposed on the carrier, with inphase (e.g., cosine) and quadrature (e.g., sine) versions of the carrier and summing the result. The signal emanating from the summing junction is the carrier properly shifted in frequency. A block diagram of this process is illustrated in Figure 6.17. A circuit with this architecture is easily implemented on a small segment of a single chip so integration with other functionality is possible. Gilbert mixers are a current popular selection for

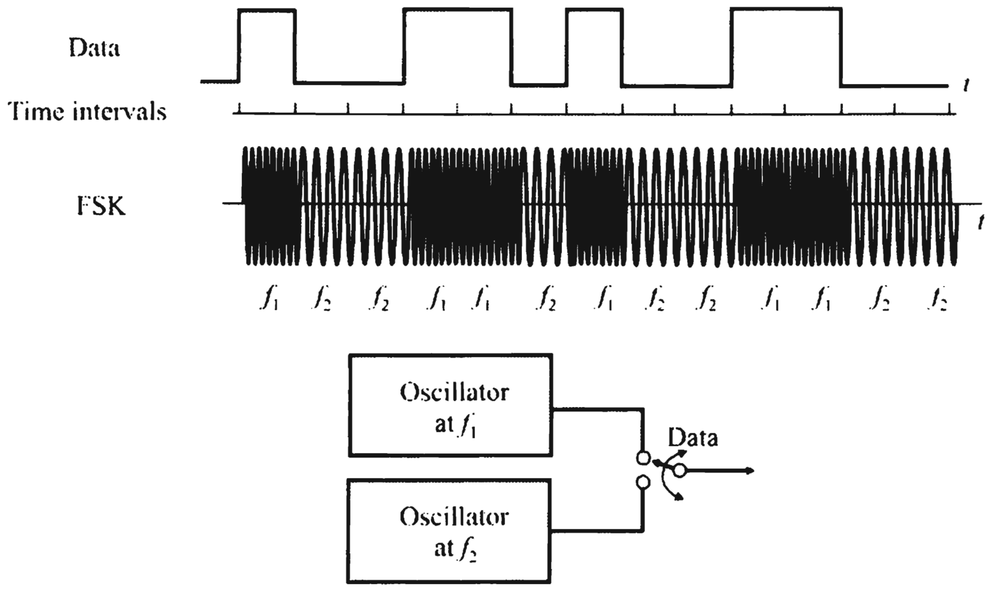

Data
Time intervals
FSK
t
t
f_1 f_2 f_2 f_1 f_1 f_2 f_1 f_2 f_2 f_1 f_1 f_2 f_2
Oscillator at f_1
Oscillator at f_2
Data

Figure 6.16 FSK changes the transmitted frequency between two (or more) frequencies.

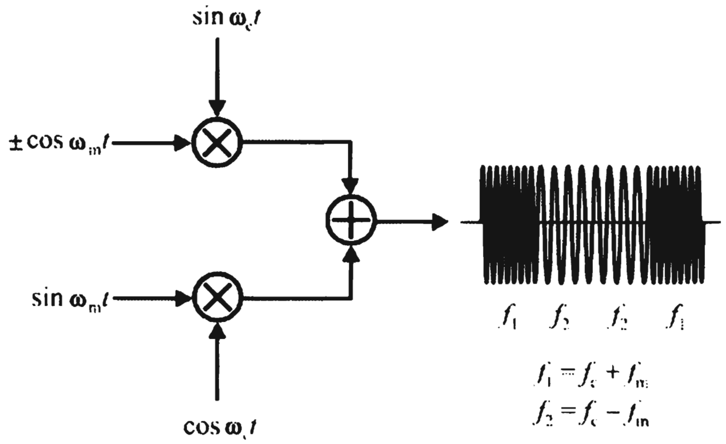


Figure 6.17 Vector modulator.

the mixers in a vector modulator. These are true four-quadrant mixers that lend themselves to chip integration.

The function of the vector modulator can be implemented in analogue or digital form. While the digital form has an upper frequency limit today (2008) of approximately 100 MHz, the analogue single chip form can be used up to about 3 GHz. At the higher frequencies, however, these chips have imperfections in the mixing and summing process. These imperfections in practice result in an image of the modulation signal appearing at the output of the modulator causing distortion in the modulated signal.

If the carrier feeding the vector modulator is at the final output frequency for the transmitter, then the modulation is imposed directly onto the output signal in what is termed a *direct conversion transmitter.*

While FSK (FM) modulation is illustrated in Figure 6.17, in fact any complex modulation can be imposed this way (amplitude, frequency, phase) onto a carrier by proper selection of the parameters.

Minimum Shift Key

Minimum shift key (MSK) is a special type of FSK where phase continuity is maintained at the symbol boundaries. It is implemented with a frequency shift that is equal to one-half of the modulating signal bit rate. This frequency shift is the smallest it can be while still maintaining orthogonality. Since phase continuity is maintained, there are no rapid changes in the phase, and MSK maintains the constant modulus (amplitude, defined below) property even after modulation. The MSK spectrum is one of the cleanest available in that little energy leakage occurs.

Gaussian MSK

When the modulating signal is passed through a filter with the transfer characteristics given by

$$H(f) = e^{-\alpha f^2} \tag{6.20}$$

prior to being modulated on carrier via MSK, it is called *Gaussian MSK* (GMSK). Constant α is a filter parameter that controls the slope of the skirts of the filter. Gaussian MSK exhibits good spectral efficiency with sharp cutoff at the band edges. This cutoff characteristic is very important for narrowband applications such as cellular radio. It also has reasonable power efficiency. It is used in second-generation cellular systems such as global system mobile (GSM) at 1.35-bps/Hz.

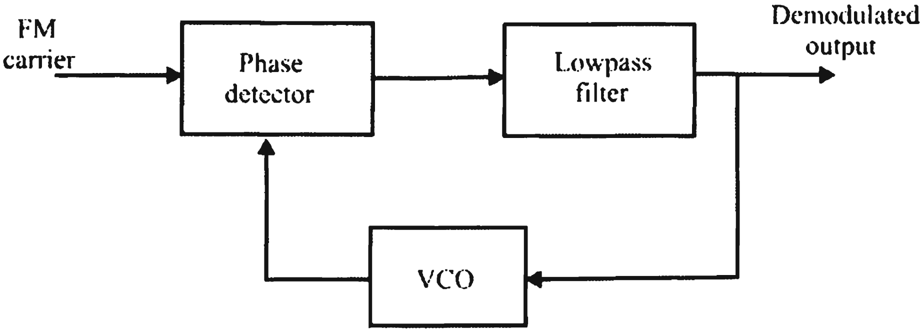


Figure 6.18 VCO in a phase-locked loop for demodulation of MSK, including GMSK and analog FM.

GMSK can be demodulated in several ways. The VCO scheme indicated in Figure 6.18 is one way. The instantaneous phase of the VCO is compared to the incoming FM signal. The output of the phase detector will increase or decrease depending on this phase difference. The increasing and decreasing function is precisely the modulated data signal. Note that this scheme also works for analog FM and regular FSK.

Alternately the simple clipper/discriminator can be used. Like most other forms of digital modulation, however, better performance can be achieved with coherent demodulation. Typically 3 dB improvement in SNR performance is achieved by using coherent demodulation.

When cellular phone networks were first implemented, they used analog modulations—specifically FM was the modulation of choice. The North American version was called the *advanced mobile phone system* (AMPS). In Europe, on the other hand, each country installing such systems used their own standard. This, as one would expect, caused these systems to not be interoperable and roaming among countries was impossible. This caused the European community to develop the *group special mobile* (GSM), a second-generation standard based on digital signaling technology. It is a TDMA-based system with TDD.

In the United States, as is common with the rest of the developed world, there are competing standards for digital cellular services, the most popular of which is IS-95, which is a CDMA modulation technique. The other standard is IS-54, which utilizes TDMA technology. Initially by edict of the FCC in the United States both of these digital standards were forced to be dual-mode, capable of both digital and analog operation. IS-136 is the same TDMA standard as IS-54 except the dual-mode operation required with older analog AMPS is removed.

There are digital data mobile services available in many areas as well as digital cellular services. Three of the most popular mobile data services are the packet switched data network, MOBITEX, cellular digital packet data (CDPD), and *advanced radio data information service* (ARDIS). MOBITEX and CDPD use GMSK, while ARDIS uses FSK. These are regional services that typically are used for business intranets, although their bandwidth/data rates are somewhat limited. Third generation (3G) cellular standard are bringing relatively high speed networking that make these services somewhat obsolete. The stated purpose of the 3G standards is to bring video and high speed internet services to the mobile environment. Downlink (telephone system to mobile user direction) data rates in excess of 2 Mbps are available with the 3G cellular standards. These 3G standards use several forms of modulation, but in order to get high speeds over radio channels usually requires some form of phase-shift key modulation.

6.2.2.8 Phase-Shift Key

The phase of the carrier signal is changed in *phase-shift key* (PSK) modulation, while the amplitude (modulus) is held constant. There are various types of PSK as discussed herein. As the types get more complex, as one would imagine, the implementation gets more complex as well. Due to the fact that the phase of the carrier must be measured, some form of synchronization of the receiver with the carrier is required. After this synchronization is obtained it must be maintained; the latter is called phase tracking.

The types of modulations that are most often used for power efficient communications are *binary PSK* (BPSK), *quadrature PSK* (QPSK), and *offset QPSK* (OQPSK). BPSK uses two phase states (or changes from one symbol period to the next), 0 and $\pm\pi$ radians (0 and $\pm 180^\circ$), to carry the information signal. It is actually the transition from one state to the other that is usually of importance. QPSK uses four states, 0, $\pm\pi/2$, and π radians. OQPSK also uses four states, but they are offset from zero typically by $\pi/4$ radians. MSK is also a popular modulation scheme. It is OQPSK with half-sinusoidal pulse shaping prior to modulation. As previously indicated, MSK has some special characteristics for spectral efficiency.

BPSK

The simplest form of PSK is BPSK. The phase of the carrier is changed between two states, corresponding to 1 and 0, or 1 and −1. The time waveform is illustrated in Figure 6.19.

Implementation can be as simple as multiplying the carrier by +1 and −1, which, mathematically, is the same as shifting the phase by π radians. Because of this simplicity, it was frequently used in earlier forms of satellite communications and spread spectrum implementations. It is not as efficient, either power or spectrally, however, as other forms of PSK discussed below.

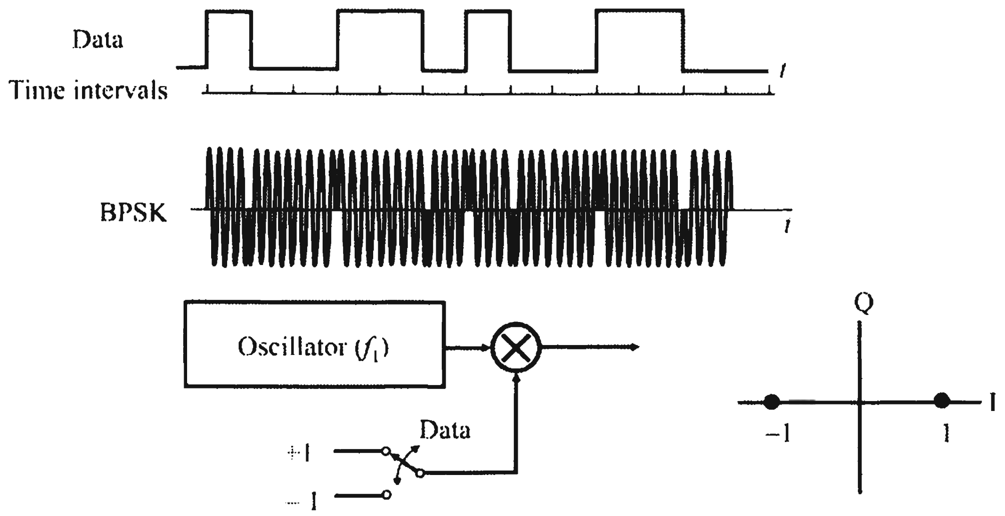


Figure 6.19 BPSK modulation.

Differential BPSK (DBPSK or DPSK) is a form of phase modulation that does not require synchronization with the local oscillator. The current symbol is based on the last symbol and, if different, a different symbol is transmitted than if it is the same.

QPSK

In QPSK, two data bits are combined together to form a single symbol as shown in Figure 6.20. As shown on the constellation, each symbol is π/2 radians apart; they all have the same amplitude (usually taken to lie on the unit circle for simplicity). Notice

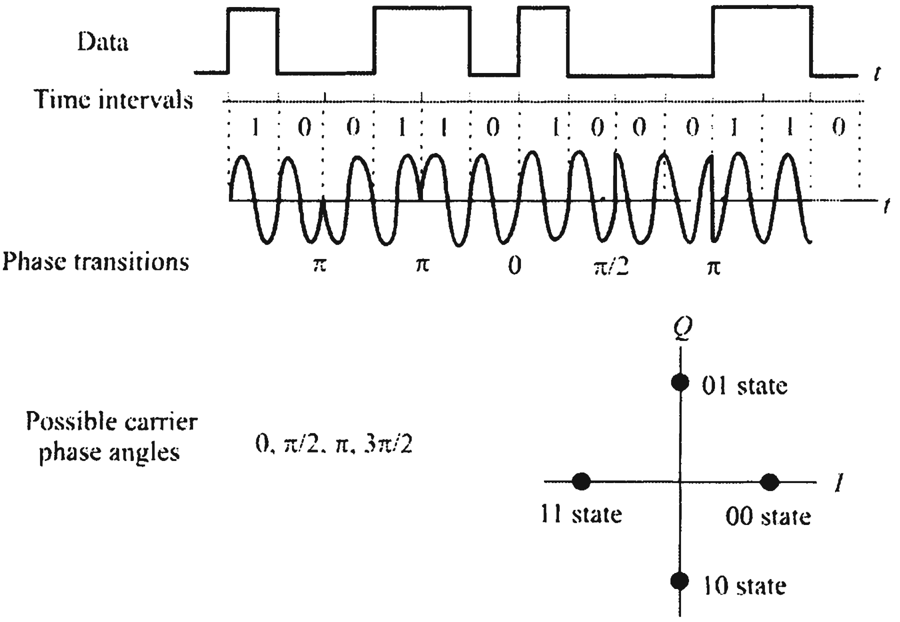


Figure 6.20 QPSK modulation. Two bits at a time are combined to form a symbol.

that there are amplitude variations of the carrier where there are phase discontinuities. The amplitude variations are greatest when the phase shifts π radians.

OQPSK

As illustrated in Figure 6.20, normal QPSK exhibits a nonconstant amplitude characteristic. This can be seen at the phase transition from 0 to π where the amplitude can go from the maximum of the carrier to the opposite extreme. Because of this amplitude variation, QPSK is not appropriate for any implementation where the amplifiers are operated close to saturation in a highly nonlinear region, such as satellite communications and digital cell phone systems. Such amplitude variations cause intermodulation products to be generated. To address this deficiency in QPSK, π/4 OQPSK was devised. The constellation for this modulation scheme is shown in Figure 6.21. Only ±π/4 and ±3π/4 phase transitions are allowed. This prevents the large amplitude variations from occurring by not allowing the signal to pass through the origin where the amplitude is zero. Upon each symbol transition, the constellation shifts counterclockwise by π/4 radians, which is where the prefix "offset" comes from. So at $t = t_i$, the symbol at π/4 might correspond to the data bits 00, while at $t = t_{i+1}$, the symbol at π/2 corresponds to the data bits 00. The whole constellation shifts so the the other symbols shift likewise.

Minimizing Intermodulation Distortion

Satellite communications are one of the chief methods for communications over long distances. The amplification device on satellites is called a traveling wave tube (TWT); it generates the power levels that allow communications at the frequencies at which satellites operate. These active devices are operated at or near saturation to achieve maximum efficiency. At saturation, the amplification becomes nonlinear, and nonlinear devices generate amplitude distortions of the signals they are amplifying. One way to minimize such distortions is to use modulation techniques that allow the modulated carrier signal to have constant amplitude (also called the modulus). Constant amplitude modulations are also required in order to reduce the sidelobes in the spectrum generated by operating the high-power amplifier (HPA) in a nonlinear mode.

Like GMSK, π/4 QPSK modulation is popular in second-generation cellular phone systems. The U.S. digital AMPS standard, IS-54, uses π/4 OQPSK at 1.62 bps/Hz. The Japanese digital cell implementation uses it as well at 1.68 bps/Hz. Some others are the European Tetra at 1.44 bps/Hz and the new Japanese personal handy phone (PHP) system.

Table 6.1 Efficiencies and Other Data on Simple Forms of Digital Modulations

Modulation	**Bandwidth Efficiency**		E_b/N_0 **(dB) at** $P_b = 10^{-5}$	**Immunity to Nonlinearity**[1]	**Implementation Complexity**[2]
	Nyquist	**Null to Null**			
BPSK	1	0.5	9.6	D	A
QPSK	2	1	9.6	C	B
OQPSK	2	1	9.6	B	C
MSK	N/A	0.67	9.6	A	D

Source: [9]. [1] Ranking from A to D (A = best, D = worst); [2] Ranking from A to D.

In summary, nonlinear amplification (on satellites, in cellular handsets, and elsewhere where power efficiency is at a premium) is the primary motivation for using $\pi/4$ OQPSK modulation, where the constellation is shifted by $\pi/4$ each successive symbol interval with the result that $\pm\,\pi$ transitions are not allowed. This allows for a more constant signal envelope, which reduces out-of-band energy, while facilitating the performance advantages of QPSK. Table 6.1 [8] shows some common modulation techniques and compares them with the above metrics.

2^nPSK

As just illustrated for QPSK, modulating bits can be combined together to form higher types of constellations for transmission. This is referred to as 2^nPSK, where implementations have included n as high as 8 and higher.

6.2.2.9 Quadrature Amplitude Modulation

When both the phase angle and amplitude are used to encode data bits into symbols, it is called quadrature amplitude modulation (QAM). The amplitude and phase of the carrier signal both depend on the value of the binary word to be encoded. The phase refers to the change of the phase of the carrier from the last symbol as opposed to an absolute value. Both I and Q amplitudes as well as phase are modulated. There are 2^n discrete symbols in QAM. When $n = 1$, BPSK ensues. When $n = 2$, it is the same as QPSK above. The inphase component is referred to as I and the other component is referred to as the quadrature component, or Q. I and Q transmit two data streams simultaneously, as discussed in Section 6.2.3.

As an example, suppose $n = 3$ corresponding to 8QAM. Further suppose that the coding process follows Table 6.2. Using Table 6.2, the following digital stream

001011010100101011101010100110101000

Table 6.2 Phase Shifts for the Example

Binary Word	Amplitude	Phase
000	1	None
001	2	None
010	1	π/2
011	2	π/2
100	1	π
101	2	π
110	1	3π/2
111	2	3π/2

separated into octets for clarity becomes

001 011 010 100 101 011 101 010 100 110 101 000

which is coded as shown in Table 6.3.

The constellation for the above 8-QAM is shown in Figure 6.22. This 8-QAM coding would facilitate 3 bits per symbol, where a symbol is one of the points in the constellation. For one symbol period, 1/(1,200) second for 1,200 baud and 1/(2,400) for 2,400 baud, for example, that symbol is transmitted. If the channel is a voice grade channel such as that used for modems over telephone lines, for example, then the resultant data rate would be 3,600 bps for 1,200 baud and 7,200 bps for 2,400 baud. The amount of encoding is a power of two, and 8-QAM, 16-QAM, 64-QAM, and 256-QAM are not uncommon. Furthermore, the symbol points need not lie on the axes.

Modern modems that are used to interconnect personal computers (PCs) to the local Internet service providers use QAM as the modulation technique below 33 Kbps. The advantage is speed. Local telephone lines only provide a bandpass of 300–3,000 Hz, typically, and at 1 bps/Hz (a speed faster than older encoding schemes) the

Table 6.3 Phase Changes

Data Word	Amplitude	Phase Change
001	2	None
011	2	π/2
010	1	π/2
100	1	π
101	2	π
011	2	π/2
101	2	π
010	1	π/2
100	1	π
110	1	3π/2
101	2	π
000	1	None

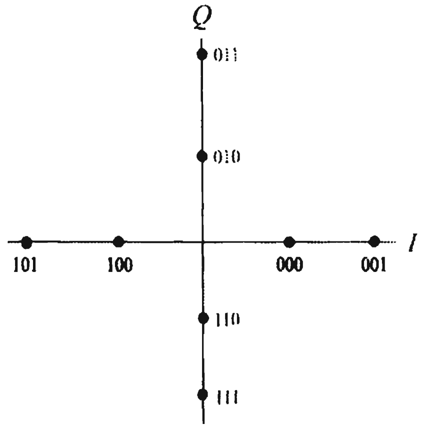


Figure 6.22 Constellation for 8-QAM.

throughput of these lines could only go to about 3,000 bps. With QAM encoding, higher than 1 bps/Hz can be achieved. Each symbol in 16-QAM, for example, represents one of 16 states, which is 2^4, so four bits are sent for each symbol, resulting in a 4:1 increase.

QAM is used extensively for terrestrial radio links. It is not suitable for satellite links, however, because of the amplitude variations inherent in the modulated signal as previously discussed.

6.2.2.10 Spectral Efficiency

Various modulation schemes exhibit different amounts of energy spillover into adjacent channels, as well as channels further away. Figure 6.23 shows a comparison of these modulation techniques. One of the major tradeoffs, as is evident from Figure 6.23, is the amount of energy in the next channel versus the amount of energy in channels further away [9].

6.2.2.11 Linear Modulations

This is a term that broadly refers to the cases where the amplitude is changed to transfer information—the information could be in analog or digital form. The general definition for linearly modulated signals is [10]:

> Let $\varphi[f(t)]$ be the modulated signal and $f(t)$ be the modulating signal. Then the modulation scheme is *linear* if $\{d\varphi[f(t)]/df(t)\}$ is independent of $f(t)$. Otherwise it is a nonlinear modulation.

Thus, AM is a form of linear modulation since

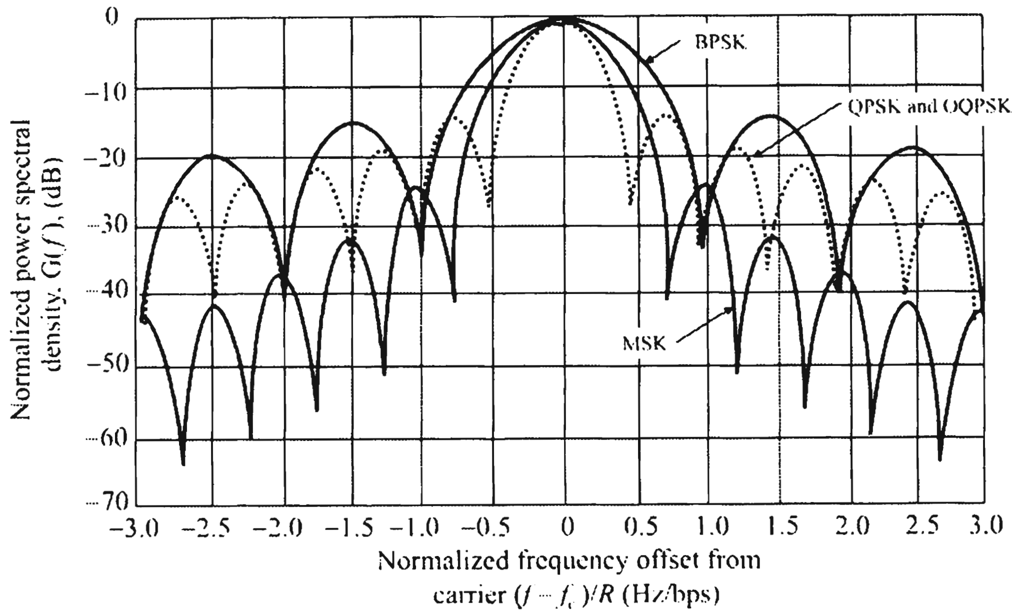


Figure 6.23 Spectral efficiency of MSK, BPSK, QPSK, and OQPSK. (*Source:* [9], © IEEE 1994. Reprinted with permission.)

$$\begin{aligned} s(t) &= f(t)\cos(2\pi f_c t + \Psi_c) \\ \frac{ds(t)}{df(t)} &= \frac{df(t)\cos(2\pi f_c t + \Psi_c)}{df(t)} \\ &= \cos(2\pi f_c t + \Psi_c) \end{aligned} \tag{6.21}$$

and $\cos(2\pi f_c t + \Psi_c)$ is independent of $f(t)$. The phase can also be modulated in such techniques.

The reason to focus on linear modulations is their relative spectral efficiency. FM can be interpreted as a form of spread spectrum modulation, although it is generally not thought of that way. Part of the definition of spread spectrum is that the spreading signal is independent of the information signal and that is not true with FM. The modulated signal is typically much wider in bandwidth than the bandwidth of the information signal, however. That is not true for AM signals. The bandwidth of AM-DSB is only twice the bandwidth of the information signal, while for AM-SSB, the RF bandwidth is the same as the information signal. Therefore, such signals have better spectral efficiency. The downside is, of course, that they are generally more susceptible to common forms of noise and interference. The latter is becoming less of a problem in PSTN applications where fiber is used as the transport medium. Fiber-optic networks are inherently much less noisy than either cable or wireless.

Such narrow bandwidths could be problematic since receiving equipment must know the frequency of operation very specifically. One way to accomplish this is to

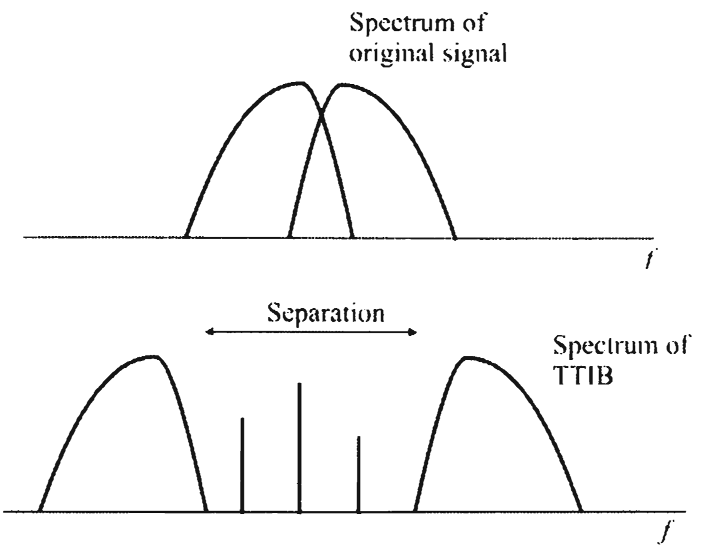


Figure 6.24 Spectrum of TTIB.

imbed a tone in the signals. This type of signaling is referred to as *transparent tone in band* (TTIB), discussed next. The tone is placed at a very specific place in the bandwidth, which is measured to tune the receiver precisely.

As the RF spectrum becomes more crowded, spectral efficiency becomes more important. Newer digital modulation schemes are packing higher bit rates into a given bandwidth—8 bps/Hz and higher are not uncommon. Channel allocations in the future in all parts of the RF spectrum can be expected to be narrower. In the military VHF range in the United States, the channel allocations are currently 25 kHz. These will soon be decreased to 12.5 kHz and 6.25 kHz. Using complex digital modulation schemes, these narrower bandwidths can carry the same information rate that is carried today over the 25 kHz channels. Channel coding is used to combat the higher noise susceptibility of linear modulations.

TTIB

It is sometimes convenient to transmit a synchronization tone with a signal. These signals can be used to synchronize a local oscillator for synchronous detection, for example. One way to do this is with a TTIB. As shown in Figure 6.24, the baseband signal is split apart (although literally the tone need not be in the center of the band) and one or more tones are inserted in the band gap. At the receiver, the tones are removed and the signal is put back together. Not only can information be transferred with the tones, but the width of the band gap can be modulated to transmit information as well.

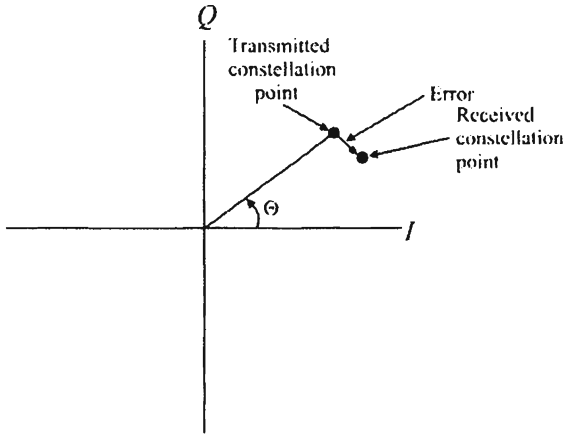


Figure 6.25 Effects of noise on QAM signals.

6.2.2.12 Effects of Noise on Digital Modulations

Either externally induced or internally induced noise, the latter at either the transmitter or the receiver, can cause a received symbol to be in error. The effects are illustrated in Figure 6.25 for QAM, but are similar for any phase-modulated signal. If the noise causes the symbol to appear closer to the expected location of one of the other symbols, the receiver will decide incorrectly which symbol was transmitted. Thus, the BER is a function of the SNR of the signal.

Figure 6.26 shows the SNR required to produce the indicated BERs for several types of popular digital modulations [11]. These SNRs are postdetection, or demodulated, SNRs, denoted SNR_d herein.

6.2.2.13 Equalization

Equalization is the process of filtering a received signal to account for imperfections in the communication channel. The channel is measured for its frequency response by the transmitter sending tones to the receiver. The receiver knows these tones and the time of their transmission. The receiver ascertains how well those tones were received and sets the filter coefficients accordingly. Adaptive equalization for digital communication channels needs to be applied. Without equalization, ISI becomes too large and intolerable interference results, essentially precluding communications. ISI occurs when one information symbol interferes with others. Whereas adaptive equalization is normally applied to wireline communication channels as well, especially for communication by modem discussed in Section 6.4, it is also important for wireless communications. For mobile data communications, an expanding area of digital communications, including the digital overlays of the cellular networks, the modem techniques developed for wireline communications are normally the techniques of choice since it is a well-developed field.

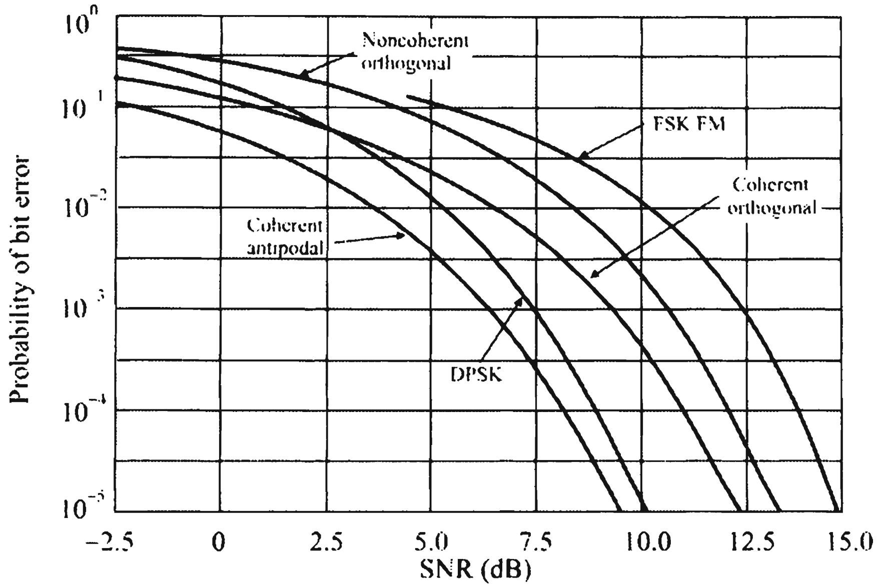


Figure 6.26 Demodulated SNRs required for the specified level of performance (P_e). (*Source:* [11], ©John Wiley & Sons, 1988. Reprinted by permission.)

Subsequent symbols in digital channels are delayed and arrive at a receiver later than the symbol that arrived via the direct route. ISI can significantly degrade the channel's performance whenever the delay time (also called the delay spread) is greater than about 10% of the symbol time.

6.3 Spread Spectrum

There is a tradeoff that can be made when transmitting information through a channel. It was discovered by Claude Shannon and is described by a theorem given by

$$C = B\log_2(1+\delta) \qquad \text{bits/sec} \qquad (6.22)$$

Equation (6.21) says that the channel capacity *C*, in symbols per second, can be increased either by increasing the bandwidth *B*, in hertz, or increasing the SNR δ. Note that this is a theoretical limit and is independent of the type of modulation and other channel parameters.

One method of increasing the bandwidth of a signal is with FM. Another way is to implement spread spectrum modulations. There are two fundamental techniques that fall within the normal definition of spread spectrum: frequency hopping and direct sequence. In the former, the carrier frequency of the signal is changed periodically, or "hopped." In the latter the energy in a relatively narrowband information signal is spread over a much larger bandwidth. Spread spectrum is implemented for a variety of reasons, in addition to increasing the channel capacity. It provides for a degree of covertness, because in frequency hopping the receiver needs to know to what frequency the signal has hopped. This information is normally not known a priori to narrowband ES systems. In direct sequence, since the energy is spread over a wide bandwidth, at any narrowband portion of the spectrum occupied by the signal there is very little energy—frequently below the level of thermal noise present. Thus, the signal is hard to detect.

Another use for spread spectrum is for range measurements. This is particularly true for direct sequence, which is the method used in GPS for determining the range from the satellites to any particular point close to the Earth. Accurate range measurements are possible because of the precise timing in GPS provided by atomic clocks. The correlation between sequences is measured very accurately and the time offsets are used to determine range differences.

6.3.1 DSSS

Consider the communication system shown in Figure 6.27. The symbol $j(t)$ represents an interfering signal, such as a jammer while $\eta(t)$ represents noise contributed by the communication channel. A data sequence $d(t) = \{d_i\}$, $d_i \in \{-1, +1\}$ is multiplied by a chip sequence $c(t) = \{c_i\}$, $c_i \in \{-1, +1\}$ as illustrated in Figure 6.28. The chip rate, $1/T_c$, is much higher than the data rate, $1/T_b$. Furthermore, the number of chips per data bit is an integer, $N = T_b/T_c$. The product $d(t)c(t)$ multiplies the carrier signal $\cos(2\pi f_0 t)$, forming the transmitted signal

$$s(t) = \sqrt{2S}d(t)c(t)\cos(2\pi f_0 t) \tag{6.23}$$

where S is the power in the signal; the energy per bit is thus $E_b = ST_b$. This is one form of a BPSK signal. Another equivalent form is given by [12]

$$s(t) = \sqrt{2S}\cos(2\pi f_0 t + d_n c_{nN+k}\frac{\pi}{2}) \tag{6.24}$$

where $k = 0, 1, 2, \ldots, N-1$, n is an integer, and $nT_b + kT_c \le t < nT_b + (k+1)T_c$. The processing gain of a spread spectrum system, denoted by G_p, is given by the ratio of the bandwidth of the spread signal to that of the nonspread signal. It represents the advantage of a spread spectrum system. Thus if the bandwidth of a nonspread signal is B_s and this signal is spread by some means to occupy a bandwidth of B_{ss}, then

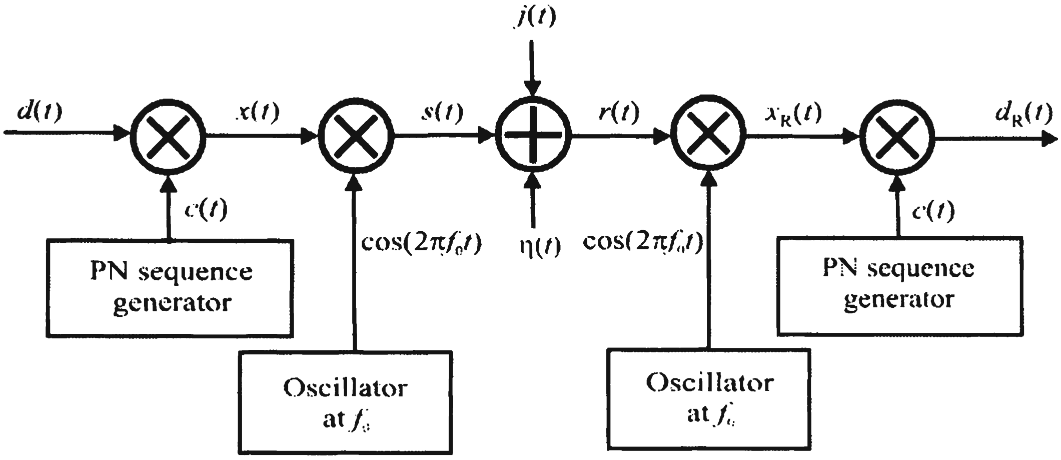


Figure 6.27 BPSK DSSS system.

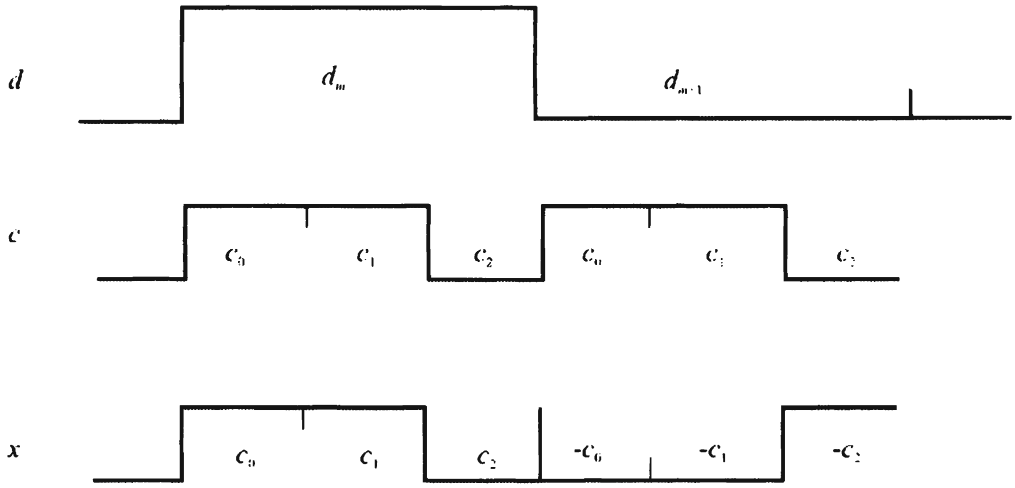


Figure 6.28 Multiplying the data sequence by the chip sequence.

$$G_{\mathrm{p}} = \frac{B_{\mathrm{ss}}}{B_{\mathrm{s}}} \tag{6.25}$$

The received signal $r(t)$ is given by

$$r(t) = s(t) + j(t) + \eta(t) = \sqrt{2S}d(t)c(t)\cos(2\pi f_0 t) + j(t) + \eta(t) \tag{6.26}$$

This signal is down converted (filters, amplifiers, and other components are left off the diagram for clarity) by multiplying $r(t)$ by a replica of the signal from the transmitter oscillator. The resulting signal contains components at $2f_0$ and around $f = 0$. It is assumed that the $2f_0$ term is filtered away, leaving only the component around $f = 0$. It is further assumed that these oscillators are at the exact same frequency and are synchronized. The signal is thus

$$x_{\mathrm{R}}(t) = \sqrt{2S}d(t)c(t)\cos(2\pi f_0 t)\cos(2\pi f_0 t) + j(t)\cos(2\pi f_0 t) + \eta(t)\cos(2\pi f_0 t) \tag{6.27}$$

Suppose $j(t) = \eta(t) = 0$ for now. Then

$$\begin{aligned} x_{\mathrm{R}}(t) &= \sqrt{2S}d(t)c(t)\cos^2(2\pi f_0 t) \\ &= \sqrt{2S}d(t)c(t)\left(\frac{1}{2} + \frac{1}{2}\cos(2\pi 2 f_0 t)\right) \\ &= \frac{1}{2}\sqrt{2S}d(t)c(t) + \frac{1}{2}\sqrt{2S}d(t)c(t)\cos(2\pi 2 f_0 t) \end{aligned} \tag{6.28}$$

The second term is filtered out, leaving only the first term.

This signal is then multiplied by $c(t)$, yielding

$$d_{\mathrm{R}}(t) = \frac{1}{2}\sqrt{2S}d(t)c(t)c(t) = \frac{1}{2}\sqrt{2S}d(t)c^2(t) \tag{6.29}$$

where it is assumed that the two PN sequence generators are synchronized. Now $c^2(t) = 1$ for all t, so

$$d_{\mathrm{R}}(t) = \frac{1}{2}\sqrt{2S}d(t) \tag{6.30}$$

which is within a constant amplitude of the original data sequence.

The question remaining is why use this modulation format at all, since simpler techniques are available to move a data sequence from one point to another. The advantage becomes clear when $j(t) \neq 0$.

It is assumed that $j(t)$ is centered at f_0, just as the signal is. (See Figure 6.28.) In this case, after multiplying by the local oscillator and filtering off the double frequency term, the remaining signal is given by

$$x_{\mathrm{R}}(t) = \frac{1}{2}\sqrt{2S}d(t)c(t) + \frac{1}{2}j(t) + \frac{1}{2}\eta(t) \tag{6.31}$$

Multiplying this signal by $c(t)$ yields

$$d_{\mathrm{R}}(t) = \frac{1}{2}\sqrt{2S}d(t) + \frac{1}{2}j(t)c(t) + \frac{1}{2}\eta(t)c(t) \tag{6.32}$$

Now, just as the spectrum of $d(t)$ was spread at the transmitter by multiplying by $c(t)$, the jammer signal is spread at the receiver by this multiplication by $c(t)$. Thus, while the multiplication collapses the desired signal to a multiple of $d(t)$, the jammer signal is spread at the receiver, so that the energy per unit bandwidth is quite low. This is illustrated in Figure 6.29. Therefore, spread spectrum exhibits an antijam characteristic. In fact, it was this property that, to a large extent, motivated its development by the military.

The noise signal, while not correlated with the local replica of the spreading sequence, does not get spread as the jammer signal does. Instead it passes through the de-spreading process fairly unscathed. It has been shown that the noise power after de-spreading is a significant fraction (over 90%, depending on the modulation involved) of the noise power that is input to thede-spreader [13].

6.3.1.1 Spreading Codes

The codes used to generate the spreading signal, $c(t)$, are selected with specific properties. In particular, the mean value of the chips is approximately zero:

$$\bar{c} = \sum_{i=0}^{N-1} c_i \approx 0 \tag{6.33}$$

In order for this to be true, the number of -1's must approximately equal the number of $+1$'s.

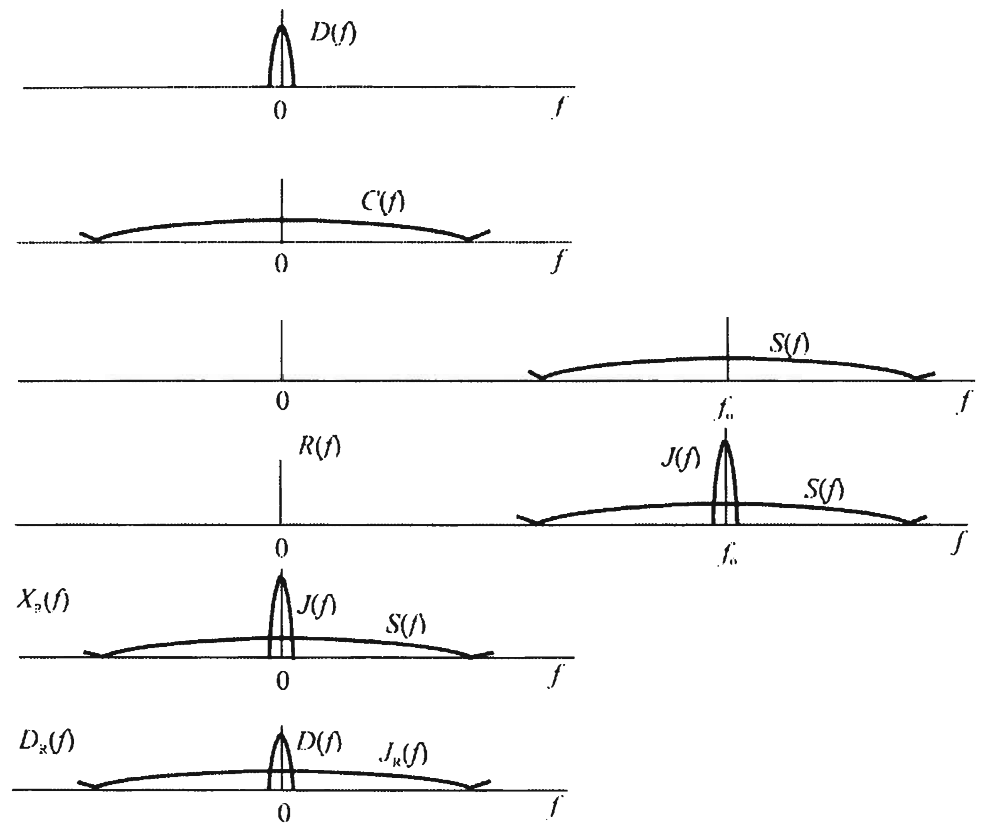


Figure 6.29 Signal spectra of the DSSS system.

The second significant property is their autocorrelation, which is approximately an impulse function:

$$R_c(k) = \sum_{i=0}^{N-1} c_i c_{i+k} \approx \begin{cases} N, & k = 0 \\ 0, & k \neq 0 \end{cases} \tag{6.34}$$

The last property that will be noted here is that their cross-correlation functions are as close to zero as possible. In spread spectrum communication systems, the signals are correlated against a locally generated sequence. If two signals correlate (their cross-correlations are nonzero) then incorrect signals could properly be demodulated at the receiver. If their cross-correlations are zero this will not occur.

In order to achieve a mean value of approximately 0, the PN sequence must have approximately the same number of -1's as 1's in any substantial length of the sequence. This difference should be 0 or only 1. This is called the balance property. The run property of a PN sequence dictates the following:

- 1/2 ($1/2^1$) of runs are of length 1;

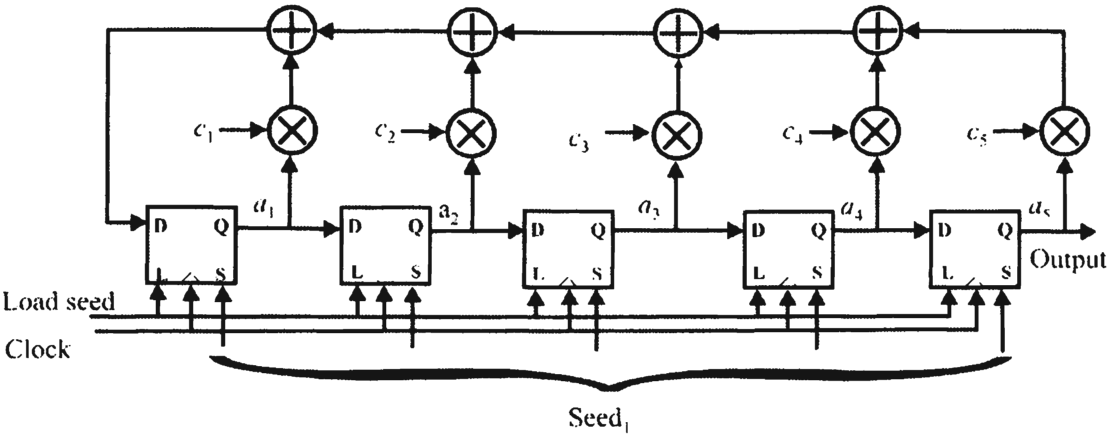


Figure 6.30 Five-stage shift register used for generating codes. The multiplying constants c_i determine whether that stage is used in the feedback loop.

- 1/4 ($1/2^2$) of runs are of length 2;
- 1/8 ($1/2^3$) of runs are of length 3.

There are three types of codes that are frequently used for spread spectrum signaling: m-sequences, Gold codes, and Kasami sequences. They all exhibit the above properties to differing extents. Codes for spread spectrum are generated with *linear feedback shift registers* (LFSRs). Associated with an LFSR is a generating polynomial that governs the code sequence generated by the LFSR. The five-stage LFSR shown in Figure 6.30, for example, has the generating polynomial

$$g(c,\ a) = 1\ +\ c_1a_1 + c_2a_2 + c_3a_3 +\ c_4a_4 + c_5a_5 \tag{6.35}$$

where all the additions are carried out modulo two, also called the exclusive OR. Also, a_i, $c_i \in \{0,\ 1\}$. The c coefficients determine whether that stage in the shift register is used in the feedback path, and therefore used in determining which code is generated. By setting $c_i = 0$, then stage i of the shift register is disabled from the feedback path, while setting $c_i = 1$ enables that stage. Since $0 + 0 = 0$, $0 + 1 = 1$, the exclusive OR simply passes the correct bit value on.

The exponents of the generating polynomial correspond to the state of the LFSR, which is nonzero. Starting from the left is stage 0, then stage 1, and so on. The first and last stages are always used while the intervening stages determine which specific code sequence is generated.

The characteristic of LFSRs that is exploited in code generation for spread spectrum communications is the output sequence that they exhibit. A register is preloaded with a seed and the clock takes the register through its sequence. Registers

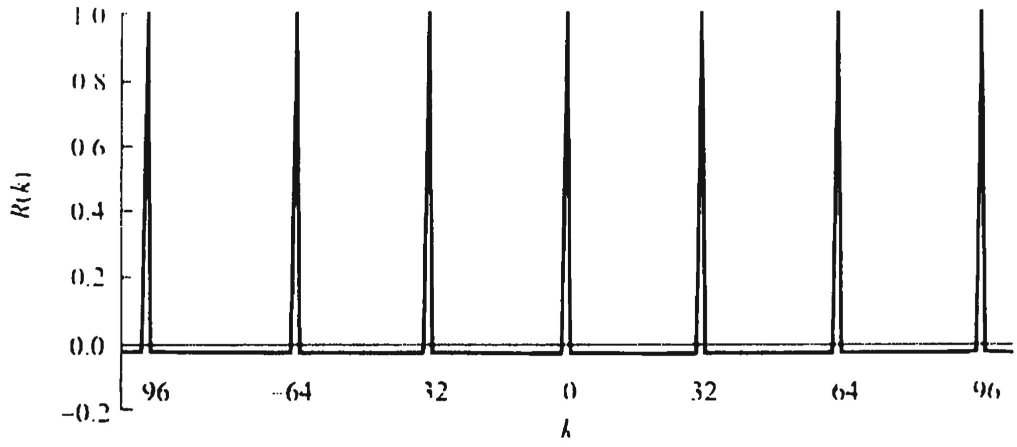


Figure 6.31 Autocorrelation function for an $r = 5$-bit m-sequence.

can be configured to generate output sequences that are as long as they possibly can be. Such sequences are called maximal length sequences, or m-sequences. For a register with r stages in it, an m-sequence has a period of $2^r - 1$. That is, the sequence repeats after $2^r - 1$ bits have been generated. The autocorrelation function of an m-sequence is periodic and is given by

$$R(k) = \begin{cases} 1, & k = nN \\ -\dfrac{1}{N}, & k \neq nN \end{cases} \tag{6.36}$$

where n is any integer and N is the period of the sequence. An example of this is shown in Figure 6.31 for an $r = 5$-bit register. Although the autocorrelation property of m-sequences is good, their cross-correlation properties are not as good as other orthogonal codes.

Using two equal length m-sequences and combining the results together generate Gold codes. An example of this is shown in Figure 6.32. Not all pairs of m-sequences can be used to generate Gold codes. Those that can are referred to as preferred pairs. The autocorrelation and cross-correlation functions of Gold codes take on three values: $\{-1, -t(m), t(m) - 2\}$ where

$$t(m) = \begin{cases} 2^{(m+1)/2} + 1, & m \text{ odd} \\ 2^{(m+2)/2} + 1, & m \text{ even} \end{cases} \tag{6.37}$$

The advantages of Gold codes are that their cross-correlation functions are uniform and bounded, and a given LFSR configuration can generate a wide family of sequences.

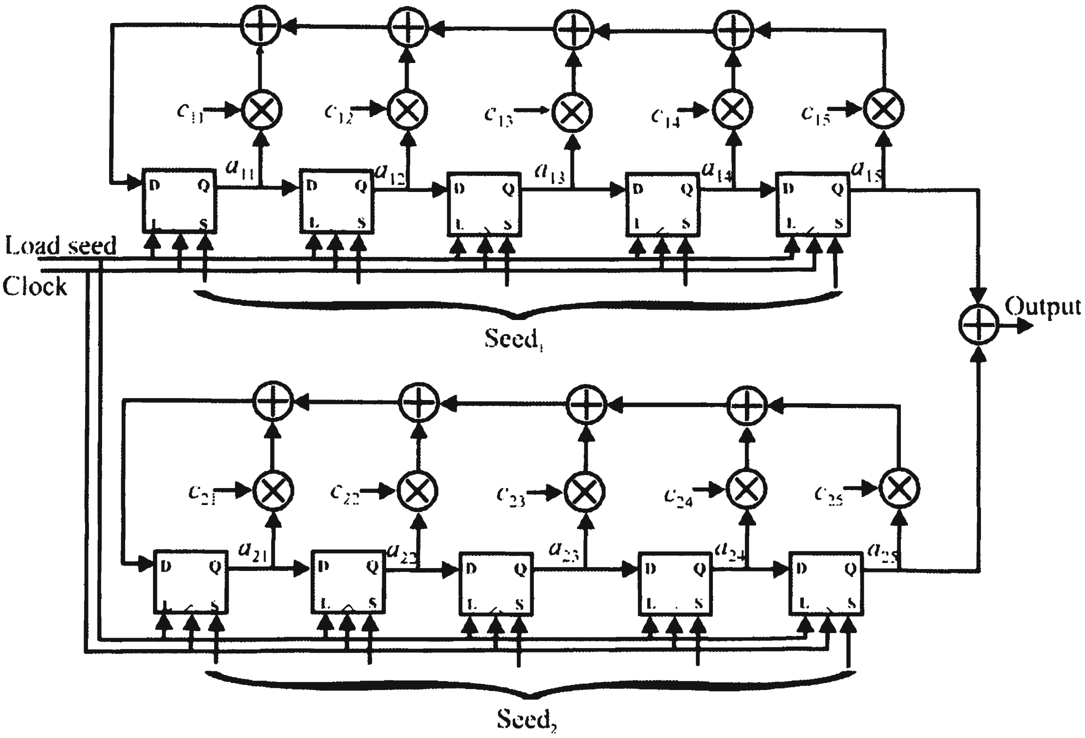


Figure 6.32 Two five-stage linear shift registers connected in parallel with their outputs added (modulo-2) are used to generate Gold codes.

Kasami sequences exhibit very low values of cross-correlation and therefore are useful in asynchronous spread spectrum systems. A similar process used to obtain Gold codes generates them, but various m-sequences are decimated and combined with the original sequence.

Walsh-Hadamard codes are orthogonal and therefore have zero cross-correlations, which makes them highly desirable for spread spectrum communications since the signals are correlated at the receiver during the reception process. The Walsh functions are simply determined from a set of matrices called Hadamard matrices, which are recursively defined as

$$\mathbf{H}_0 = [0]\ ,\ \mathbf{H}_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix},\ \mathbf{H}_4 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix},$$

$$\mathbf{H}_{2N} = \begin{bmatrix} \mathbf{H}_N & \mathbf{H}_N \\ \mathbf{H}_N & \bar{\mathbf{H}}_N \end{bmatrix} \tag{6.38}$$

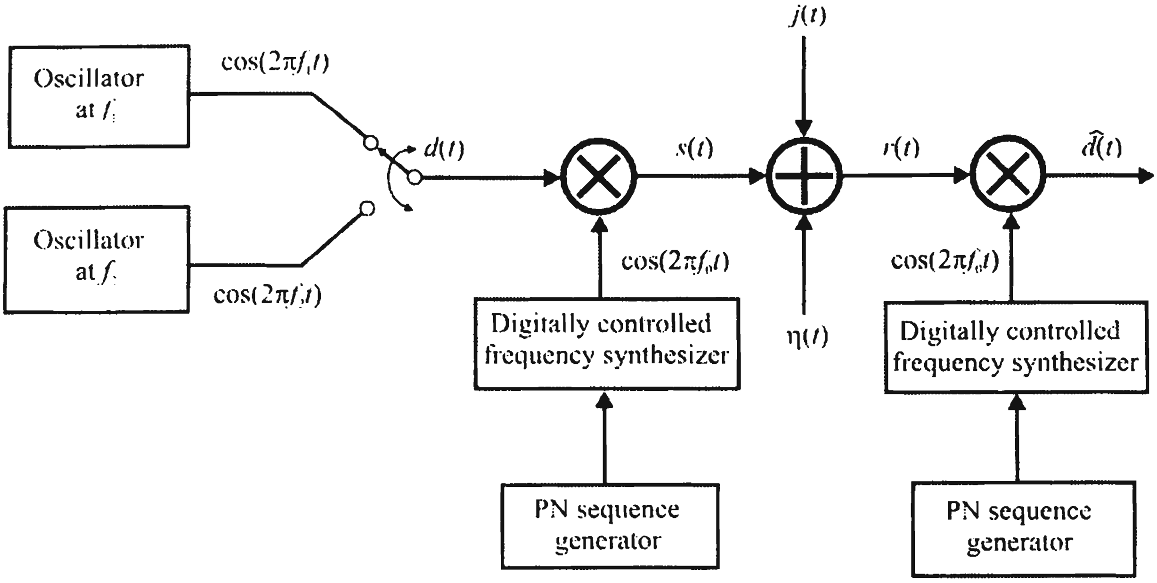


Figure 6.33 Block diagram of an FHSS communication system.

where $\bar{\mathbf{H}}_N$ is the element-by-element binary inverse of $\mathbf{H}_N$. All but the all-zero row of a Hadamard matrix forms an orthogonal code word and can be used as a generating polynomial for an LFSR. Each user in the communication system is given a different one of these code sequences and interference is minimized. Actually the all-zero row is orthogonal to the others, too, but it does not generate an interesting output.

The significant downside for Walsh functions in the spread spectrum role is that, while the whole code words are orthogonal, partial code words are not. In fact, partial code words can exhibit significant cross-correlation. In those cases where the CDMA radios are not synchronized with each other, partial correlations will occur and remove much of the advantages of orthogonal codes. For example, while the forward channel (base station to mobile handsets) in IS-95 is synchronous, the reverse channels are not so they would exhibit excessive partial correlations. For that reason Walsh codes are used in IS-95 for the forward channels but not the reverse channels.

6.3.2 FHSS

Consider the spread spectrum communication system shown in Figure 6.33. At the transmitter, the modulating signal is imposed by some modulation method—typically FM—on a carrier frequency that is varied on a regular, frequent basis. Again, amplifiers, filters, and other components are omitted from the diagram for clarity. A digitally controlled frequency synthesizer generates the carrier frequency. The PN sequence generator determines the particular frequency at any moment. This is known as an FHSS system.

The information signal $d(t)$ can be digital or analog but here it is shown as BFSK. BFSK is a popular form of modulation for frequency hopping. Although for FSK the two complementary frequencies f_1 and f_2 need not be close together, typically in the VHF frequency range they are two frequencies in the same 25-kHz channel. In many realistic applications of FHSS, noncoherent detection is employed for simplicity and economics. Therefore, even though the PN sequences must be synchronized for proper operation, the local oscillators are not phase-synchronized. Like other noncoherent communication systems, this typically results in about a 3 dB loss in performance.

At the receiver there is an equivalent frequency generator controlled by an equivalent PN sequence generator. When synchronized, the PN sequence generators change frequency control words simultaneously and in synchronization with each other. Synchronizing the PN sequence generators is typically accomplished by using a subset of the available frequency channels to which all of the communication devices tune when synchronization is required (such as when a new member joins the communication net). Synchronization information is always available on these channels.

The functions $j(t)$ and $J(f)$ represent an interfering signal, such as a jammer. The advantage of FHSS systems in the presence of jammers and other interference is illustrated in Figure 6.34. Here BFSK signals are transmitted and the frequency of transmission is varied at the hop rate. For illustrative purposes a narrowband jammer is shown, although jammer waveforms targeted against frequency hoppers would typically cover many more than just the one channel shown here. The hopping signal can hop into the jammer for the duration of the hop. In that case the information during that hop would typically be lost to the receiver. Normally coding, interleaving, and other forms of redundancy are built into the signaling scheme so that if the hopping signal hops into the jamming signal, the data is not lost. Of course, if the signal is digitized voice, some tolerance for lost data bits is inherent anyway. In any case, since the jammer is fixed in frequency, any lost information is only lost for the duration of that hop, and for prior and subsequent hop intervals, the information is transferred unimpeded.

If there is more than one data bit sent per hop frequency, then the system is said to be slow frequency hopping. On the other hand, if a single data bit is sent over several hops in sequence, then the system is a fast frequency hopping system. The hop rate is denoted as R_h and the data rate is denoted as R_d. Therefore, in a slow frequency hopping system $T_b = 1/R_d << T_h = 1/R_h$. R_h for a slow frequency hopping system of 100 hops per second (hps) with a data rate, R_d, of 20 kbps are typical values. So in this case $T_h = 10$ ms and $T_b = 50$ μs. This data rate would correspond to an approximate rate for digitized speech. On the other hand, for a fast frequency hopping system, for this same data rate, the hop rate might be 40 khps or faster.

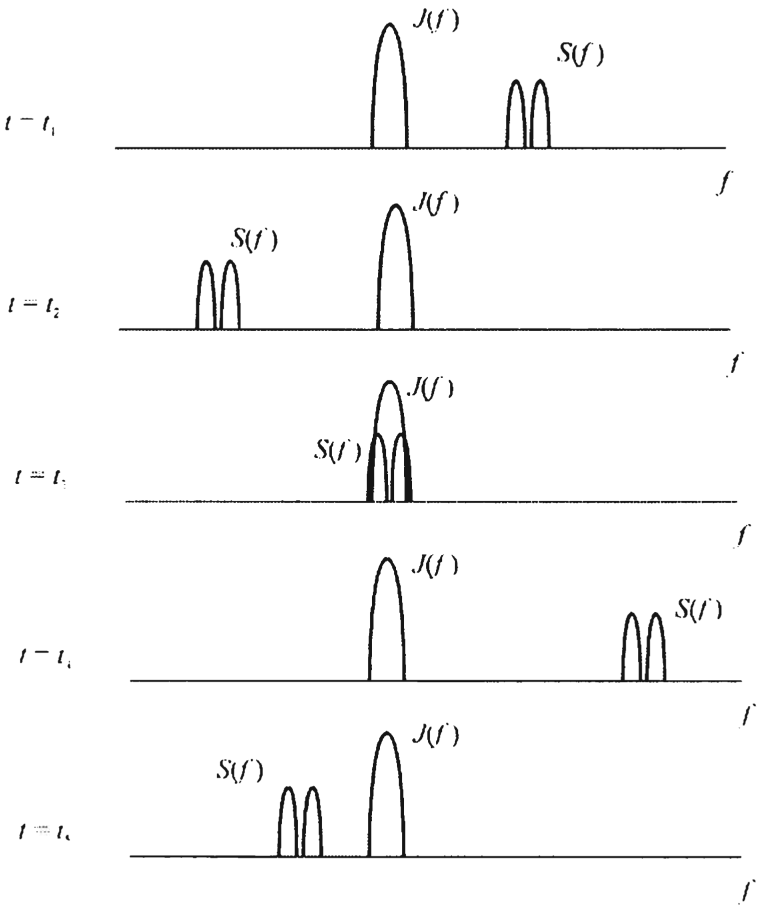


Figure 6.34 Spectra of a frequency hopping target with a fixed partial-band jamming signal.

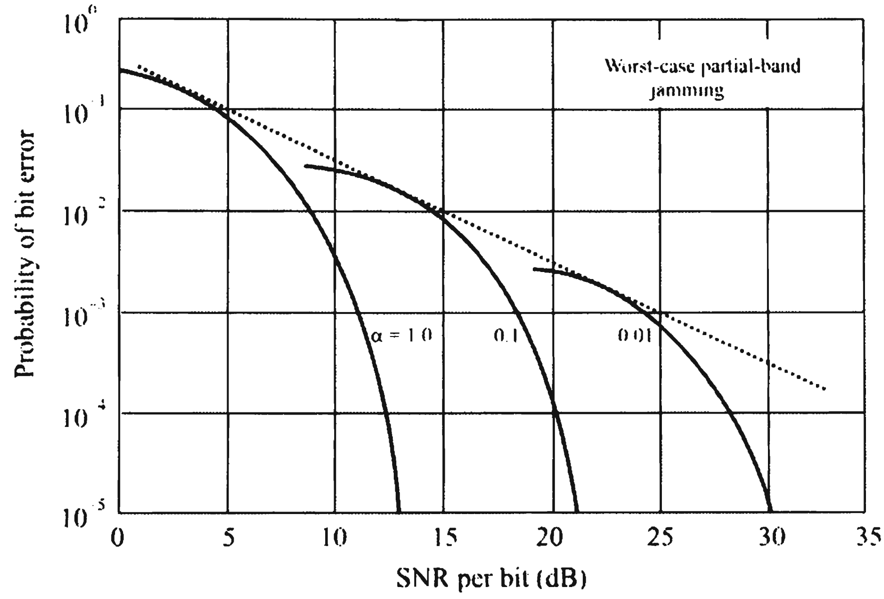


Figure 6.35 Performance of a BFSK/FHSS system in the presence of partial-band jamming. (*Source:* [14], © John Wiley & Sons 1885. Reprinted with permission.)

Let the fraction of the total bandwidth, W_{ss}, of an FHSS system that is occupied by a partial-band jammer be denoted as α. Furthermore, let γ_b denote the energy per bit, E_b, to jammer spectral density, J_0, ratio. Thus,

$$\gamma_b = \frac{E_b}{J_0} = \frac{W_{ss} / R_b}{J_{av} / P_{av}} \tag{6.39}$$

where $E_b = P_{av}/R_b$ and $J_0 = J_{av}/W_{ss}$. For a slow frequency, incoherent BFSK/FHSS system, the performance in the presence of a jammer will be as shown in Figure 6.35 [14]. As a specific example, an airborne, partial-band jammer with an ERP of 1 kW (60 dBm) is standing off a distance of 10 km from a VHF target receiver. The scenario is illustrated in Figure 6.36. Thus, the jammer waveform arrives at the receiver attenuated by approximately 80 dB, yielding $J_{av} = -20$ dBm → $J_{av} = 10^{-5}$ W at the receiver. Suppose $W_{ss} = 10$ MHz; thus, $J_0 = 10^{-12}$ W/Hz. Suppose the communication link is air-to-ground, with the receiver on the ground, and is 1-km-long. Suppose the target transmitter emits 2 W (33 dBm). Over this range in the VHF frequency range, the free-space path loss is about 60 dB, so $P_{av} = -27$ dBm → $P_{av} = 10^{-5.7}$ W. Suppose $R_b = 20$ kbps. Thus, $E_b = 10^{-5.7} \div 2\times10^4 = 0.5\times10^{-9.7}$ W·sec and $\gamma_b = 0.5\times10^{-9.7}/10^{-12} = 0.5\times10^{2.3}$ → 18 dB. The probability of bit error in this case is approximately 2×10^{-3}.

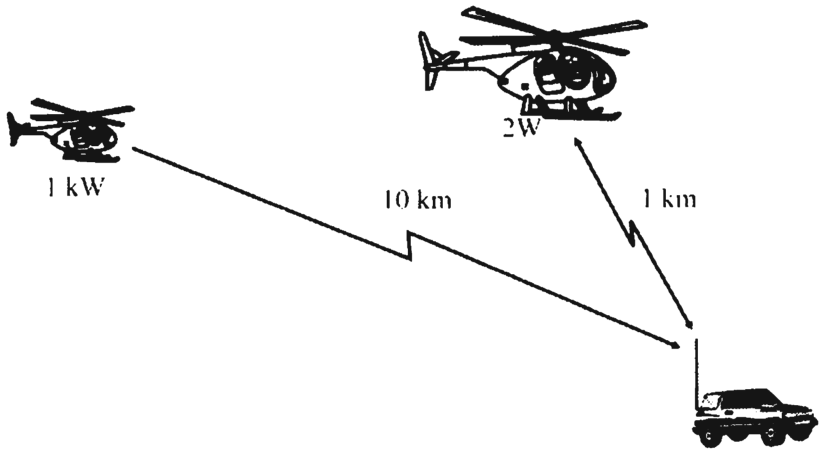


Figure 6.36 Example scenario for illustrating the effectiveness of jamming.

Code selection for frequency-hopping communication systems is similar to that described above for direct sequence. The codes should have minimum crosscorrelation and well-behaved autocorrelations. The receive process, however, is different in frequency hopping systems and, therefore, the effects of high values of crosscorrelations are different. In DS systems, when two codes had a high crosscorrelation, the improper signal could be properly decorrelated at the receiver, in effect causing unwanted interference. In FH systems, two codes with a high crosscorrelation will cause the same frequencies to occur in the two sequences more than desired. Thus, the two transmitters will hop to the same frequency more often than otherwise. This also results in unwanted interference, called cochannel interference, even though the details of the cause are different.

To combat these types of problems in spread spectrum systems the underlying data sequence is typically encoded with forward error correction data bits and the resulting data stream is interleaved. More will be said about forward error correction a little later. Bit interleaving is a process of mixing the data bits so that a burst of errors does not destroy a complete data word.

6.4 Access Methods

The techniques used to allow a user of a communication channel to use the channel is referred to as the access method of the communication system. There are principally four such access methods and they will be introduced here. Combinations of these access schemes are frequently implemented.

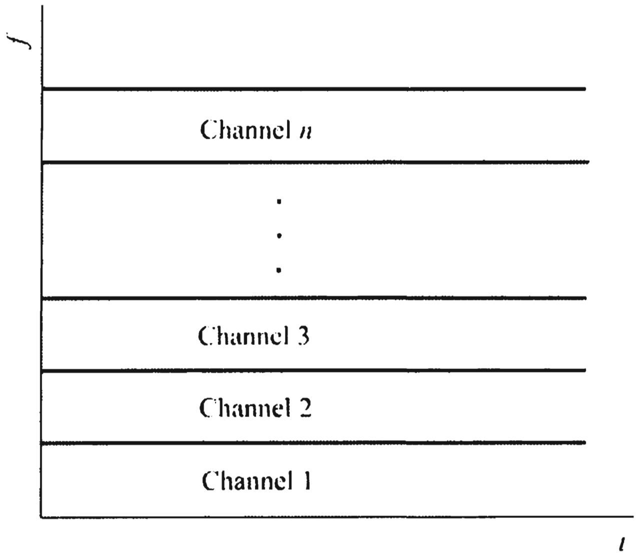


Figure 6.37 In FDMA, the frequency spectrum is divided and a user has access to that frequency channel all the time..

6.4.1 Frequency Division Multiple Access

In *frequency division multiple access* (FDMA) the available frequency channels are divided among the users. It is one of the oldest forms of access. Each user has sole use of the channel until it is no longer needed and it is relinquished. FDMA is illustrated notionally in Figure 6.37. Each user does not normally keep the channel forever, as would be implied by Figure 6.37. They do keep it whether they have anything to send or not, however, so it can be an expensive method of access.

6.4.2 Time Division Multiple Access

If a frequency channel is to be shared among more than one user at (essentially) the same time, then some method must be devised that allows this sharing. In *time division multiple access (*TDMA), use time in a channel is divided among the users. At one instant one user would be using the channel and at the next, another user has the channel. Usually this happens so fast that the users do not know that the channel is shared. TDMA is notionally illustrated in Figure 6.38. Normally the frequency spectrum would be channelized as well.

6.4.2.1 ALOHA

ALOHA is a TDMA scheme where a user transmits when desired. This results in collisions when more than one user transmits at the same time. If these collisions cause reception problems, the data will be received in error. This, in turn, causes no

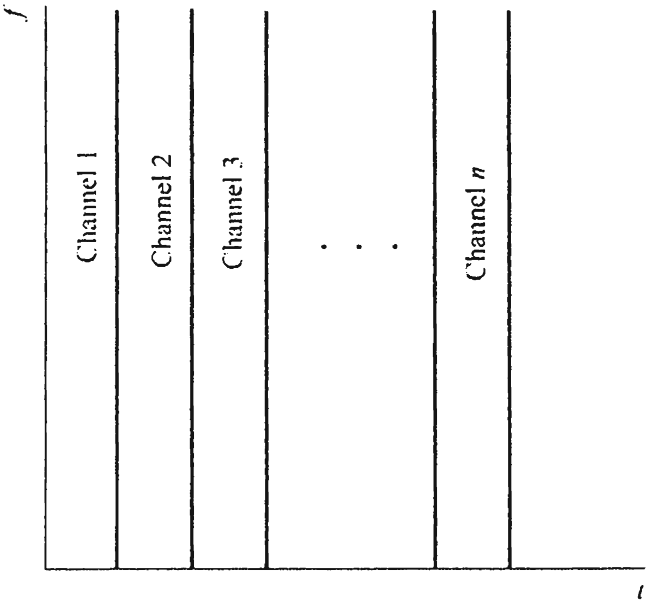


Figure 6.38 In TDMA the frequencies are shared among multiple users who each get a time slot. Normally the frequency spectrum is channelized as well.

ACK to be sent back to the transmitter or a NACK to be sent. If there is no ACK from the receiver, the message is sent again.

If

- T = Throughput measured in successful packet transmissions per frame time;
- P = Packets offered for transmission per packet time;

then the throughput of pure ALOHA is given by

$$T = Pe^{-2P} \tag{6.40}$$

As shown in Figure 6.39, the peak throughput of pure ALOHA is 0.18. That means that if the channel rate is 1 Mbps, the maximum throughput through that channel is 180 Kbps.

6.4.2.2 Slotted ALOHA

This access method, also a TDMA scheme, is similar to ALOHA except that the transmission boundaries are synchronized in time. Using the notation above, the throughput of slotted ALOHA is given by

$$T = Pe^{-P} \tag{6.41}$$

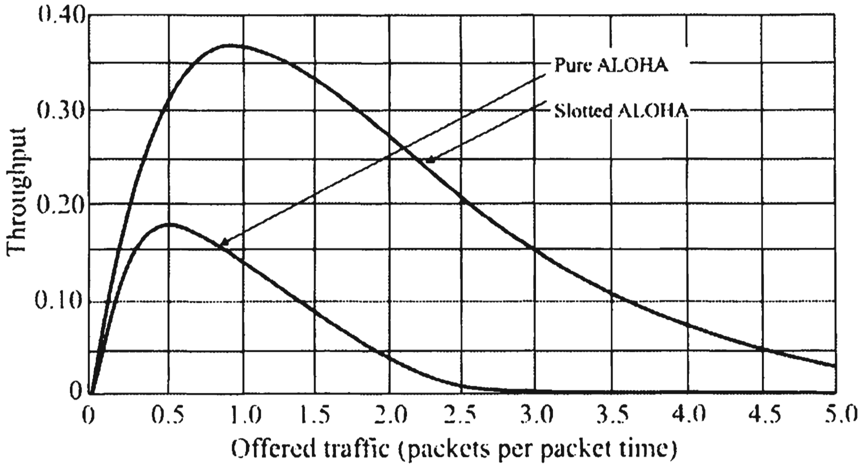


Figure 6.39 Performance comparison of ALOHA and slotted ALOHA protocols.

Thus the maximum throughput of slotted ALOHA is twice that of pure ALOHA. Its maximum throughput is 0.36 as shown in Figure 6.39, yielding 360 Kbps maximum (on average) through a 1 Mbps channel.

Although the throughput through a channel can be greater for slotted versus pure ALOHA, as shown in Figure 6.39, there is a penalty. This penalty takes the form of system complexity because of the requirement to synchronize all transmitters in the network with each other.

6.4.3 Carrier-Sensed Multiple Access

When the channel is first sensed to see if someone is already using it before a transmission is tried, it is called *carrier-sensed multiple access* (CSMA), which is a form of TDMA. It is possible to detect when a collision occurs, for example, by listening for an acknowledgment. With such collision detection the notation is CSMA/CD.

In wired networks, the maximum slotted ALOHA throughput is 36%, whereas the maximum for CSMA is 50%. Thus, there is considerable overhead with using these schemes. For wireless systems, the capture effect can increase the throughput.

6.4.4 Demand Assignment Multiple Access

A separate, lower data rate, and therefore a less expensive channel, can be used to request the use of a higher speed, and therefore a more expensive channel. Requesters are assigned usage of the channel according to some pre-established prioritization scheme. Users are notified when the channel is theirs to use. This is referred to as

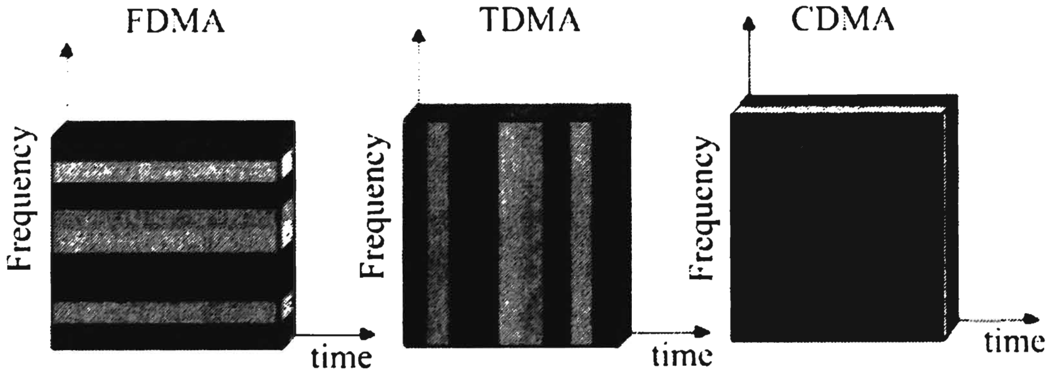


Figure 6.40 Channel access methods.

demand assignment multiple access (DAMA), and is really a form of TDMA since the channels are being shared on a time basis.

6.4.5 Code Division Multiple Access

In *code division multiple access* (CDMA), DSSS or FHSS is used to share the channel [15]. All frequencies (within a band) at all times (that have been allocated) are shared by all users. A different spreading code for each user in the band is used.

A summary of the major access methods discussed above is illustrated in Figure 6.40.

6.5 Duplexing

Duplexing refers to the method that allows a communication system to facilitate two (or more) users access to the same communication medium. In half-duplex systems, only one user can transmit at the same time without interference. Older forms of military tactical communications facilitated by HF and VHF voice radio are examples of half-duplex systems. In full-duplex systems, either or both of the users of the communication path can transmit at the same time without interfering with one another. Modern mobile digital communication systems, as exemplified by cellular phones, are full-duplex systems. There are two generally used forms of duplexing: time division and frequency division.

6.5.1 Time Division Duplexing

In *time division duplexing* (TDD) the same frequency channel is used for the upstream and downstream data but at different times (see Figure 6.41). It is conceptually similar

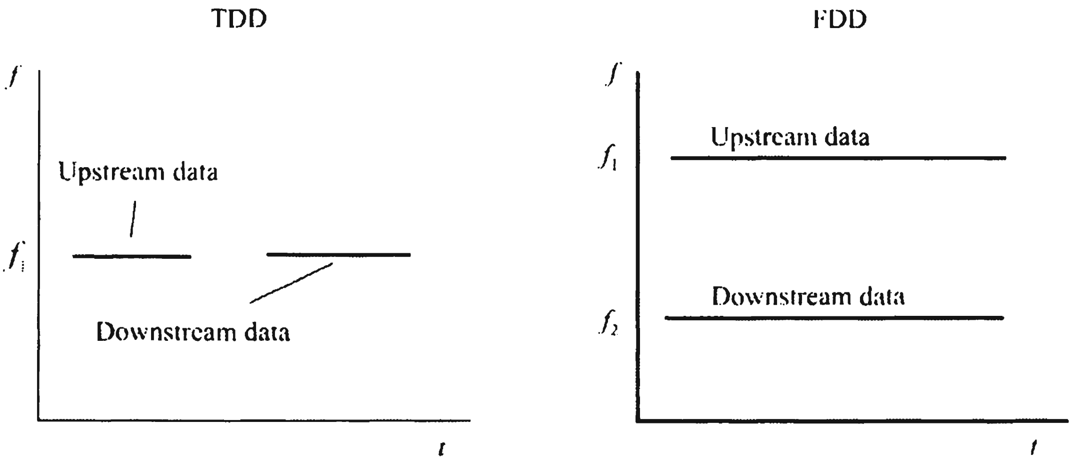


Figure 6.41 TDD and FDD are two forms of access for full-duplex communications.

to TDMA described above, but the context within which the phrase is used implies something different. Some form of duplexing is required for any full-duplex communication system, and when the two channels are separated in time but use the same frequency if is referred to as TDD.

The upstream and downstream data are sequenced in time, one after the other. There is generally a guard time allocated between the two time events in order to ensure the signals do not interfere with one another and to allow for system delays, such as those associated with finite propagation times.

6.5.2 Frequency Division Duplexing

Similarly, in *frequency division duplexing* (FDD), the data streams are sent at the same time but in different frequency channels (see Figure 6.40). This is similar to FDMA, with the difference focusing on the duplexing aspects.

6.6 Orthogonal Frequency Division Multiplexing

Orthogonal frequency division multiplexing (OFDM), also called *discrete multitone* (DMT), is a technique for mitigating the effects of poor channels. Multipath effects, interference, and impulsive noise, for example, characterize mobile communication channels. These effects can preclude high-speed communication (several megabits per second). In normal FDM systems, adjacent channels are typically separated by guard spaces—unused frequencies intended to prevent the channels from interfering with each other. This, obviously, is inefficient in terms of spectral utility.

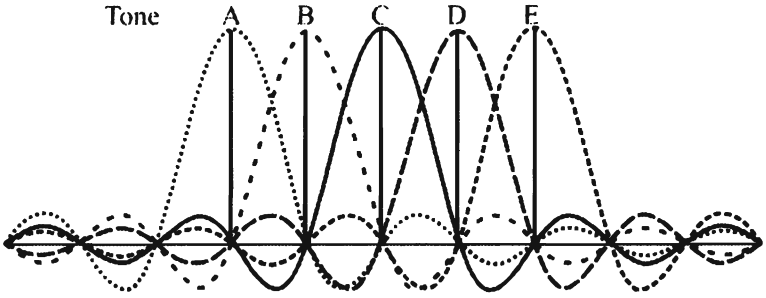


Figure 6.42 OFDM channels {A, B, C, D, E} each carry a relatively slow waveform but the signaling timing is such that they do not interfere with each other..

An alternative to this is to make the adjacent channels orthogonal to one another. Each channel then carries a relatively low bit rate, and the channel sidebands can overlap and effective signaling can still occur.

Although in a simplistic implementation, the number of carrier generators and coherent demodulators could get quite large; in fact, when viewed in the Fourier transform domain, the signal can be demodulated by taking the transform at the receiver. A completely digital implementation of the receiver is then possible by taking the Fourier transform of the received signal, as the signals at the transmitter are combined by taking the inverse Fourier transform.

OFDM is a particular type of FDM that can be viewed as a combination of modulation and multiple-access scheme. Whereas TDMA segments according to time, and CDMA segments according to spreading codes, like regular FDM, OFDM segments according to frequency as well. The spectrum is divided into a number of equally spaced bandwidths the centers of which contain carrier tones. A portion of each user's information is carried on each tone Whereas regular FDM typically requires there to be frequency guard bands between the carrier bands so that mutual interference is minimized, OFDM allows the spectrum of each tone to overlap. The tones in OFDM are orthogonal with each other, and because of this orthogonality, they do not interfere with each other.

Figure 6.42 illustrates the spectral properties of an OFDM system with five channels. This frequency domain representation of five tones in this case highlights the orthogonal nature of the tones. The peak of each tone corresponds to a zero level, or null, of every other tone. The result of this is that there is no interference between tones. When the receiver samples at the tone frequency at the center of the channel, the only energy present is that of the desired signal, plus of course, whatever other noise happens to be in the channel.

Any of the standard techniques for modulation (AM, FM, PM) can be used for modulating the information onto the tones; however phase modulation is normally used and BPSK and QPSK are typically employed. The original data stream splits into N parallel data streams, each at a rate $1/N$ of the original rate. Each stream is then mapped to a tone at a unique frequency and combined together using the *inverse fast*

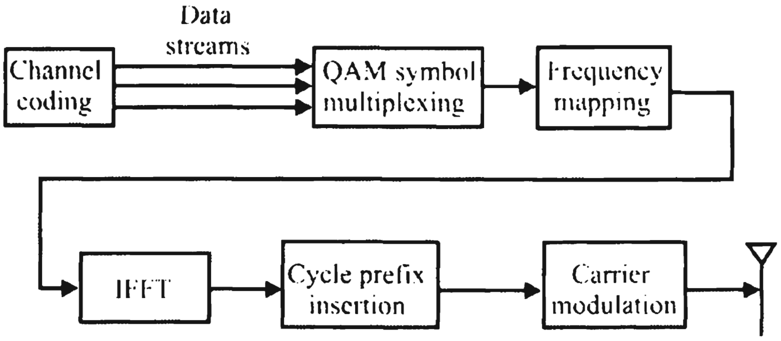


Figure 6.43 OFDM transmitter.

Fourier transform (IFFT) to yield the time-domain waveform to be transmitted. A block diagram indicating the significant functions in an OFDM transmitter is illustrated in Figure 6.43. The cycle prefix insertion function is required to provide a guard time between symbols.

As a specific example, suppose there are 100 tones, then a single data stream with a rate of 10 Mbps would be converted into 100 streams of 100 kbps each. By creating slower parallel data streams, the bandwidth of the modulation symbol is effectively decreased by a factor of 100, or, due to the inverse relationship between bandwidth and symbol duration, the duration of the modulation symbol is increased by a factor of 100. ISI can be essentially eliminated, because typical multipath delay spread represents a much smaller proportion of the lengthened symbol time. This results in the elimination of complex multi-tap time-domain equalizers at the receiver. OFDM is also a multiple-access technique. This occurs when an individual tone or groups of tones are assigned to different users. The resultant structure is called orthogonal frequency division multiple access OFDMA.

To ensure orthogonality between tones the symbol time must contain one or multiple cycles of each sinusoidal tone waveform. This is ensured by making the period of the tone frequencies integer multiples of the symbol period where the tone spacing is $1/T$ as shown in Figure 6.42. Figure 6.44 shows an example of three tones over a single symbol period, where each tone has an integer number of cycles during the symbol period.

6.7 Coding Communication Signals

Coding of communication signals is generally implemented for one of two purposes. The first is to remove redundancy in the source information. Many sources exhibit redundant characteristics, for example, speech. Coding can remove some of this redundancy to improve efficiency. So the first coding removes information. The

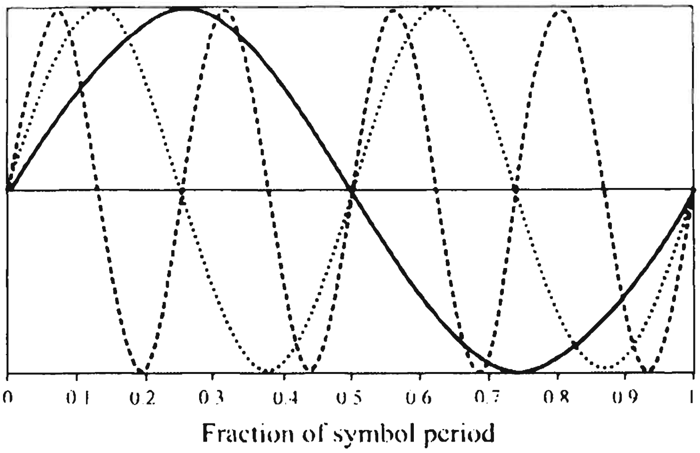


Figure 6.44 The number of sinusoid periods are integer multiples of each other. In this case there are three waveforms in the same period.

second reason for coding is to improve the reliability of the communication. Adding information to the communicated message can increase the probability that the correct message is received. So the second reason for coding adds information back.

6.7.1 Source Coding for Data Compression

Many sources of information that are common to humans contain redundant information. Human speech is a particularly bad offender in this regard. In order to maximize the efficiency of using communication resources, it may be desirable to eliminate as much of this redundancy as possible before transmitting the information. The same can be said about storing information, incidentally. The techniques for reducing this redundancy are called schemes for data compression, and some of them are discussed here. Modern day information compression techniques can facilitate the transmission of images at a 300:1 or more compression ratio in some cases without sacrificing image quality [16].

6.7.1.1 Speech Coding Techniques

An order 0 model or encoding scheme uses only the current value of the data to be encoded. An example is "u" in English. There is 1/100 chance of this letter occurring in normal English. The entropy of this character results in 6.6 bits to encode. Entropy is a term that refers to the amount of randomness in a piece of data. It is a term borrowed from thermodynamics, where it refers to the amount of randomness in gasses. The higher the randomness, the less probable the character is, the higher the entropy is, and the more bits are taken to encode it. Taking this in reverse, the more probable a character is, in order to minimize the number of bits it takes to transmit such common characters, fewer bits are assigned.

For an order 1 model, the previous character is included in the encoding process. For example, if a q is the previous character, then the probability that the current character is a u is 95%. In this case it takes only 0.074 bits (on average) to encode the current character. Source coding technology is important to understand for EW system design because source coding is used to get as much information as possible through narrow transmission channels.

Run-length encoding, used extensively in facsimile communications, encodes strings of the same character (such as blank characters) as the character followed by the number of times it is repeated. Huffman encoding allocates the number of bits representing a character in an alphabet (such as the English language) according to the number of times it appears in a message. Those characters that appear the most receive the fewest number of bits allocated, thus minimizing the number of bits necessary to send the message. Arithmetic coding accomplishes similar objectives using similar notions. There is considerable redundancy in higher levels of communications, such as English text. The Trie-based codes, LZ-77 and LZ-78, address this by using pointers to previously sent phrases, indicating where the phrase began and how long it was. All of these coding techniques are lossless. They can typically achieve a compression ratio in the range of 10:1.

6.7.1.2 Compression Techniques for Images

In both the commercial world as well as that of the military forces worldwide, the requirement to view and move images around is becoming widespread. In the commercial world, still pictures over the Internet and motion pictures over cable TV are two mainstream applications. A TV picture is comprised of 6.5 MHz of analog bandwidth. If that signal is digitized at the minimum rate, the Nyquist rate, of 9 MHz with 10 bit samples, for example, the resulting bit stream would be 90 Mbps. This is the data rate required to move this signal around as well as to store it on DVD disks. Fortunately, there is considerable redundancy in most video signals and source coding techniques are applied to reduce or eliminate that redundancy. The fundamental tradeoff in implementing these techniques is picture quality versus implementation complexity. The redundancy in images can be represented by the amount of correlation there is between the basic data in the image.

Transforms are frequently applied to reduce the transmission and storage requirement for images. Perhaps the most common one is the discrete cosine transform (DCT), but others that have been used include the Harr transform and the Walsh-Hadamard transform. The most efficient transform in a mean-squared sense is the *Karhunen-Loeve transform* (KLT); however, it is extremely complex to implement compared to some others. The DCT is popular because it is simple and its performance is close to that of the KLT.

The redundancy in video signals comes in two forms—the spatial redundancy within a single image and the temporal redundancy from one image to the next. To address the former, the Joint Pictures Expert Group (JPEG) established a standard for video compression for still pictures. The DCT is used to transform 8×8–pixel segments of the picture. The DCT coefficients are normally largest around zero and typically fall off rather rapidly. Therefore, only a few need to be retained to represent the image. In addition, Huffman encoding of the coefficients is applied which assigns shorter code words to the most frequently occurring coefficients, and run-length encoding is applied to those results. The results are extremely efficient coding of single images. Such coding of single images is referred to as intraframe coding. This is the basis for the JPG image coding popular on PCs.

To remove the redundancy from one image to the next, interframe coding is used. Just as there is considerable redundancy in the spatial context within a single image, there is significant redundancy from one image to the next. At the speeds of motion pictures, objects typically don't move much from one image to the next.

The H.261 standard uses transmission rates that are multiples of 64 Kbps so compatibility with the *integrated services digital network* (ISDN) could be maintained. It matches a 16×16 block from one image to the next and the difference is encoded as opposed to the image itself. Data rates of up to 1.5 Mbps can be implemented with this protocol.

The *motion pictures expert group* (MPEG) developed three standards for transmission and storage of motion video called MPEG 1, MPEG 2, and MPEG 6. MPEG 1 was developed to support video on CD-ROMs. Data rates of up to about 1.5 Mbps are supported as part of the standard. Forward prediction only as well as both forward and backward (bidirectional) prediction for motion compensation are part of the standard. Forward prediction is similar to that described for H.262, where motion prediction is computed. Backward prediction implements a similar concept except the current image uses the next image to reduce redundancy.

MPEG 1 allows data rates up to about 1.5 Mbps, so a higher-speed standard was needed, particularly for transmission of video over cable TV systems. Thus, the MPEG 2 standard was developed. MPEG 2 allows transmission up to about 15 Mbps. As part of the MPEG 2 standard, it is possible to specify SNR requirements and the coding adapts to those requirements. Thus, it is possible to specify the quality of the signal ahead of time so the proper coding can be applied and the images are coded accordingly. The 90-Mbps TV data rate mentioned above can be encoded with MPEG 2 at around 5 Mbps.

Lossy coding can be applied in cases where a certain amount of loss is allowed in coding an image. A still picture, for example, can be adequately represented with some loss in the data in the picture. Lossless coding is when losses are not desired or allowed. X-ray coding might be an example of this, where it is critical to transmit an image with the maximum resolution possible. In lossless situations, the investment in moving or storing images is less important than keeping all the data possible.

At the other end of the speed spectrum, MPEG 4 was developed. It was intended to apply to videophone applications, such as videoconferencing, where rapid movement was not likely, some loss could be tolerated in the transmission, and the regular narrowband PSTN is the transport medium (at 2,700-Hz guaranteed bandwidth). This standard implements speeds of 5 to 64 Kbps.

6.7.1.3 Chrominance Subsampling

All of the above coding techniques for compressing video images implement what is called chrominance subsampling. Each pel is represented by three 8-bit values—one for luminance, or brightness, and two for chrominance, or color. If two adjacent horizontal pels are the same color, determined by averaging the original colors and transmitting or storing just the one color value, then it is called 4:2:2 subsampling. This coding reduces the data rate requirements by one-third and there is little perceptual difference in the image.

Likewise, down-sampling the colors can be applied in the vertical direction in the image. This is called 4:2:0 subsampling, but normally the changes in the image are detectable. The compression rate of one-half is accomplished this way.

If the information in the image can be transferred adequately in black and white, then, of course, all of the color information can be removed. Videoconferencing, for example, may not need the color information to be effective. Transmitting black and white documents also does not require color. The result is a black, white, and gray image. The data rate in this case has been reduced by two-thirds, however.

6.7.1.4 Steganography

Steganography is the technique of hiding a message within another message such that detection of the presence of the first message is prevented. One way to accomplish this is to embed messages in image files. Such files are typically comprised of many pixels and therefore bytes. Depending on the information message to be sent, it may be necessary to change only a few such bits. The image within which the message is embedded is called the cover image. An algorithm of some sort inserts the information message into the cover image. The intended receiver of the message knows both the original cover image as well as the coding algorithm so the information-bearing message can be recovered. The encoding algorithm could take many forms. Calculating the exclusive OR on a few of the pixel bits would be one way. Changing 1 bit in every pixel or every *n* pixels would be another.

Images are either color or gray scale, and either can be used for such purposes. The more random the image, the easier it is to hide the presence of the information message. Changing the least significant bit of the pixel information, for example, one that describes a color image changes the color only slightly—typically imperceptible to the human eye. If a pixel is represented by 3 bytes, or 24 bits, changing 1 bit per pixel in a large image would go unnoticed.

6.7.1.5 Communication EW Implications

Except for MPEG 4, the video standards described here are too wideband to be of much concern to forward-deployed tactical forces. The data rates involved in transporting H.261, MPEG 1, and MPEG 2 preclude the use of most tactical radios. The bandwidth requirements dictate that the higher frequency ranges are required, where adequate bandwidth is available. Communication devices at the division echelon and above during conflict and at all levels otherwise have adequate capacity to carry such communications. Tactical radio systems that are typical of forward deployments, in general, do not have an adequate capacity. The MPEG 4 standard can be used with these radios to transport video in the near term

This is not to say, however, that the higher data rate signals are not of concern. Video teleconferencing has become a popular way for military forces to coordinate and plan their activities, and therefore the communications over which this transpires is of considerable interest. Systems and devices to attack such communications are much different from those for tactical communications, however. In those cases where wired and/or fiber cables can be utilized, bandwidth is usually not aproblem.

Steganography presents a particularly challenging problem for EW systems. A human observing an intercepted image will rarely be able to tell that a message is hidden in the image. A machine will only be able to tell the presence of the message if it has access to the original cover image and the encoding algorithm, under which circumstances the EW system would not be required.

6.7.2 Channel Coding for Error Control

Channel coding is normally performed in communication and data storage systems (which in reality are a form of communication system where the goal is communication in the fourth dimension—time) for error control. No realistic communication channel is noise-free—noise in this case being defined as anything other than the intended signal. Because of this bit errors inevitably occur.

Some types of communications will tolerate bit errors better than others. Digitized speech, for example, will tolerate a considerable level of bit errors before the speech becomes unintelligible. On the other hand, computer-to-computer digital exchanges, such as bank transactions, must be accomplished almost error-free. In the latter case, great lengths are taken to insure that information is exchanged correctly.

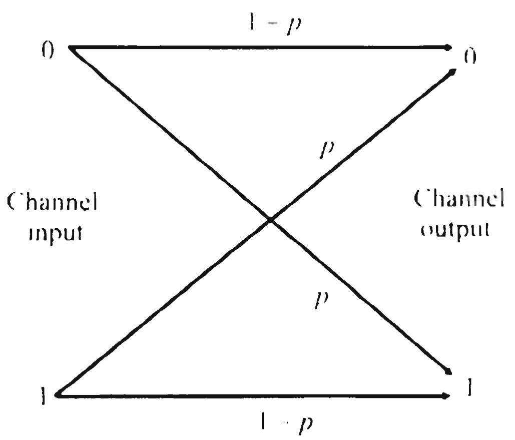


Figure 6.45 Graphic depiction of a binary symmetric channel.

Thus, coding is used to perform this error control. In some cases simply detecting errors is sufficient. In that case either the data within which the bits in error is simply discarded—possible in speech, for example, or the data can be discarded and the source can be requested to resend the data. In other cases it is possible to encode data so that, up to some limit, errors in the data can be corrected at the receiver. There are tradeoffs with both approaches.

One of the most significant advantages of digital communications is the ability to detect and/or correct errors in the data streams. This is done using coding techniques. Although a complete discussion of error detecting and correcting coding is beyond the scope herein, some of the more important topics for IW will be discussed.

The channels over which communication takes place are herein assumed to be *binary symmetric channels* (BSCs). What this means is that the probability of an error occurring is independent of the bit that was transmitted. If the probability of no error occurring is denoted by p, also called the reliability of the channel, then the probability of an error is given by $1 - p$. This is represented by the graph in Figure 6.45. Note that if there is a modem involved (discussed later), it is normally considered as part of the channel.

6.7.2.1 Error Detection

It is possible in some system implementations to allow the loss of some data but it is undesirable to let data in error persist. Also, some implementations perform error correction on large portions of a data stream as opposed to the individual pieces of data in it, but the way the accuracy of these large portions is checked is to check each piece of data for correctness. These are examples of when error detection can be used.

Table 6.4 Parity Bit Attached to Data Words

Data Word	Parity Bit	Bit in Error		
		1	**2**	**3**
00	0	100	010	001
01	1	111	010	010
10	1	001	111	100
11	0	010	100	111

Parity

One or more bits can be added to a data word to facilitate error detection. Perhaps the simplest of these schemes is adding a parity bit. A bit can be added so that the total number of 1s in a word (a word is defined here as any set of bits greater than two) is even, in which case it is called even parity, or odd, in which case it is called odd parity. As an example, suppose there are four words to be sent, given by 00, 01, 10, and 11. Further suppose that it is known (to a reasonable degree of assurance) that errors can only occur in one bit in any word. Then these words, encoded with even parity, are given in Table 6.4. The data words to be encoded are listed in the left column, and the even parity bit is given in the second column. The possible received encoded words consist of the correct encoded word, or one of those in the last three columns. By inspection it can be seen that none of the encoded words received with errors in them are legitimate so an error occurred in the transmission and it can be detected.

Digital data can be organized into a matrix, such as that shown in Table 6.5 (even parity is used). These 8-bit bytes might represent data bytes from some file to be transferred. A parity bit can be attached for each row and one for each column, shown as the bottom row and rightmost column. If a single error occurs, it will not only be detected but it can also be corrected since the row and column in error will identify a specific bit.

Cyclic Redundancy Check

The cyclic redundancy check (CRC) is a very popular coding technique for error detection and is frequently used in automatic repeat request (ARQ) schemes. It is popular because it is easy to implement, it has very good error detection capabilities, and there is little extra that needs to be done to check for errors. There are integrated circuits available that compute the CRC.

It is based on the notion that binary data streams can be treated as binary polynomials, and it is possible to add, subtract, multiply, and divide these streams of

Table 6.5 Examples of Adding (Even) Parity Bits to Rows and Columns

1	0	0	1	1	0	0	1	**0**
0	1	0	1	1	0	0	1	**0**
0	1	1	1	1	1	0	0	**1**
1	0	0	1	0	0	0	0	**0**
0	1	1	1	0	1	1	1	**0**
0	0	1	0	1	1	0	0	**1**
1	0	0	0	0	0	1	0	**0**
0	0	0	0	1	1	1	0	**1**
1	**1**	**1**	**1**	**1**	**0**	**1**	**1**	

data just as if they were numbers. The binary numbers are the coefficients of the polynomial. For example, if the binary word is given by 1010011, then the corresponding binary polynomial is given by

$$g(x) = 1x^6 + 0x^5 + 1x^4 + 0x^3 + 0x^2 + 1x^1 + 1x^0 \tag{6.42}$$

or simply,

$$g(x) = x^6 + x^4 + x + 1 \tag{6.43}$$

The data stream that CRC operates on is called a frame and a *frame check sequence* (FCS) is calculated for every frame. In CRC, the FCS is calculated so that the data sequence resulting from concatenating the original data frame with the FCS is exactly (no remainder) divisible by a polynomial. The particular polynomial is selected depending on the type of errors expected to be encountered. Some common polynomials are

$$\begin{aligned}
\text{CRC}-12: g(x) &= x^{12} + x^{11} + x^3 + x^2 + x + 1 \\
\text{CRC}-16: g(x) &= x^{16} + x^{15} + x^2 + 1 \\
\text{CRC}-\text{ITU}: g(x) &= x^{16} + x^{12} + x^5 + 1 \qquad (6.44) \\
\text{CRC}-32: g(x) &= x^{32} + x^{26} + x^{23} + x^{22} + x^{16} + x^{12} + x^{11} + x^{10} \\
&\quad + x^8 + x^7 + x^6 + x^4 + x^2 + x + 1
\end{aligned}$$

If $d(x)$ is the binary polynomial corresponding to the data frame to be encoded, then first $r(x)$ is obtained by dividing $d(x)$ by $g(x)$:

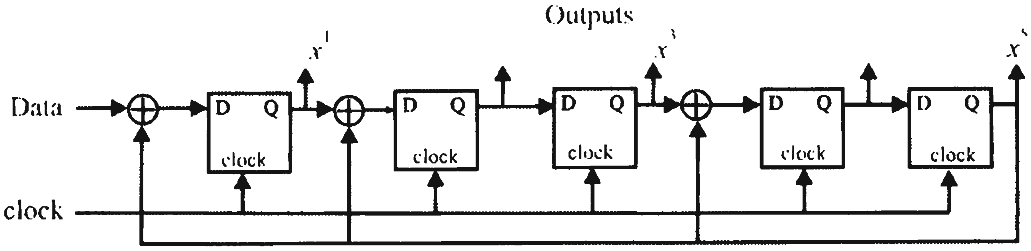


Figure 6.46 Linear shift register for generating and computing a CRC, in this case based on the polynomial $g(x) = x^5 + x^3 + x + 1$.

$$\frac{d(x)}{g(x)} = q(x) + r(x) \tag{6.45}$$

The resultant data frame that is sent is then the original frame with the FCS appended to it. At the receiver, this sequence is divided by $g(x)$ and if there were no errors in the transmission, then the remainder of this division is zero.

This code detects error bursts of length less than or equal to $(n - k)$ with probability 1, bursts of length equal to $(n - k + 1)$ with probability $(1 - 2^{-(n-k-1)})$, and bursts of length greater than $(n - k + 1)$ with probability $(1 - 2^{-(n-k)})$ [17]. It has been estimated that when using a 17 bit polynomial, the probability of allowing a 1-bit error go undetected is one in $10^{16.}$

Implementation of the CRC algorithm is easy in either hardware or software. The hardware implementation using shift registers is shown in Figure 6.46, where the polynomial $x^5 + x^3 + x + 1$ is implemented. The exclusive OR gates (the circles with the + in them) correspond 1 to 1 with the existence of a nonzero coefficient in the generating polynomial, with the exception of the highest-order coefficient. These gates have as their inputs the prior stage and the highest-order bit as feedback. To use this circuit, when generating the FCS, the data frame is first shifted into the shift register from the left with n zeros appended to the end. After the register has been clocked $n + k$ times, it will be holding the FCS. To check a received frame it is shifted into the register from the left. After $n + k$ clock cycles, the register will be holding all zeros if the frame is error-free.

The CRC procedure works because of the properties of prime numbers. A prime number is a number that can only be divided by itself and one, without generating a remainder. Within certain reasonable assumptions, if an integer of a given number of digits is divided by a prime number then the remainder is unique among all of the numbers with the same number of digits. The shift register in Figure 6.46, for example, effectively divides a data stream by the prime number 41 $(2^5 + 2^3 + 2^1 + 2^0)$.

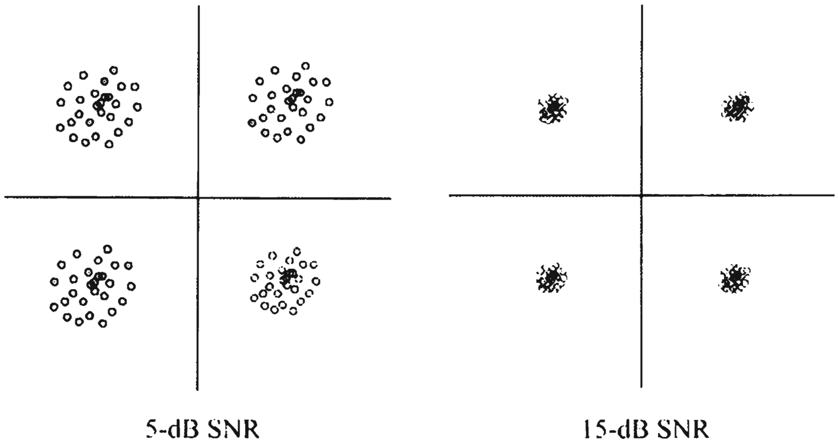


Figure 6.47 Effects of noise on a QAM constellation.

Checksum

A checksum is computed for a set of data words by summing the values of data, without worrying about overflow, as if the data were always numbers. Thus, the addition is performed modulo 2^n, where n is the number of data bits in a word. The checksum is computed on the data set to be sent (or stored) and is attached to the data as if it were part of the data set. When the data is checked, the checksum is added along with the data and the result is zero unless there is an error. It accomplishes a reasonable degree of error detection. If data is completely random and each data word is 16 bits long, then there is about a 1 in 64,000 chance that the checksum would be computed accurately for an inaccurate data set.

6.7.2.2 Error Correction

There are two generally accepted techniques for error correction: *forward error correction* (FEC) and *automatic repeat request* (ARQ). In the former, bits are added to the outgoing data stream so that an algorithm at the receiver can correct bits that are received in error. In ARQ, extra bits are also added to the outgoing data stream, but at the receiver, the received data stream is checked for errors, and if any are present, a request to resend the data is made.

A constellation for OQPSK, shown in Figure 6.47, exhibits some variations in the location of the data points. In the figure, a constellation with a low SNR is shown along with one with a relatively higher SNR. At a sufficiently low SNR, the receiver will declare some of the symbols wrong. Furthermore, all communication systems are limited in signal power at some level. The lower this level is, the closer together the symbols are in the constellation. Therefore, many digital communication systems employ error control, which can be either error detection or error correction.

ARQ

ARQ is a relatively simple form of error correction, but it is also one of the least efficient. Using one of the error detection techniques described above or others, a received data word is checked for correctness. If it is determined that the data is in error, a NACK is sent back to the transmitter indicating that the data should be sent again. When it is resent, the whole data sequence is resent. Thus in low SNR environments, the channels can quickly get clogged resending messages received in error. FEC schemes were devised to address this inefficiency issue inherent with ARQ.

FEC

When bits are added appropriately to each digital word prior to conversion to symbols in order to correct errors, it is called FEC. It does, of course, reduce the number of data bits that can be sent in a given bandwidth but in many cases this is a favorable tradeoff. Of course, as the bandwidth is increased, more data bits can be sent but this means that more noise enters the receiver, thereby increasing the BER. The goal, then, in FEC code design is to the reach an optimum tradeoff between the number of bits added versus the bandwidth necessary to transmit the code words.

There are three major types of error correcting coding techniques: (1) block codes, (2) cyclic codes, and (3) convolutional codes. Block coding usually encodes large blocks of data at a time, and the encoding depends only on the current data to be encoded. Convolutional coding encodes a block of data, and the encoding depends on the current block of data to be encoded, as well as prior blocks of data. To use the best features of both block coding and convolutional coding, they are frequently combined.

For a simple example of error correction, suppose that there are only two possible data words that can be sent given by 01 and 10. Furthermore, suppose that it is known (to a reasonable degree of assurance) that only one bit error can occur in any symbol. One possible encoding is shown in Table 6.6 for this case. The data word to be sent is in the leftmost column and the error bits attached are in the second column. These two columns concatenated could be received when there is no transmission error. The

Table 6.6 Attaching Error Correction Bits to Data Words

Data Word	Error Correction Bits	Bit in Error			
		1	2	3	4
01	00	1100	0000	0110	0101
10	11	0011	1111	1001	1010

remainder of the columns contain possible received encoded words depending on which encoded word was sent and which bit is in error. Comparing the two rows clearly shows that whichever encoded symbol is received it can be mapped back into the correct data word, under the assumption that only 1 bit (or none) is in error, thus forming an error correcting code.

This example points out the overhead necessary to implement error correction; 2 bits were added to correct 2-bit data words. This is an extreme case, however, and as the size of the data word increases, the number of check bits relative to the number of data bits decreases.

Burst error protection. Suppose each column of the following matrix is a data word to be sent and they are encoded so that up to 2 bits in error in each data word would be corrected at the receiver. Transmission of these data words would normally be $x_{1,1}$, $x_{1,2}$, $x_{1,3}$, ..., that is, vertically down the column. Suppose the order were changed so that the data bits were sent horizontally, that is, $\mathrm{x}_{1,1}$, $\mathrm{x}_{2,1}$, $\mathrm{x}_{3,1}$, ... Then a burst of errors up to 14 bits long could occur and the data words would be delivered correctly, because a burst this long or shorter guarantees that no more than 2 bits in any original data word are in error. Normally a burst of 14 bits would destroy two or three data words (in this case word = byte), depending on wherethe burst started.

$$
\begin{array}{ccccccc}
x_{1,1} & x_{2,1} & x_{3,1} & x_{4,1} & x_{5,1} & x_{6,1} & x_{7,1} \\
x_{1,2} & x_{2,2} & x_{3,2} & x_{4,2} & x_{5,2} & x_{6,2} & x_{7,2} \\
x_{1,3} & x_{2,3} & x_{3,3} & x_{4,3} & x_{5,3} & x_{6,3} & x_{7,3} \\
x_{1,4} & x_{2,4} & x_{3,4} & x_{4,4} & x_{5,4} & x_{6,4} & x_{7,4} \\
x_{1,5} & x_{2,5} & x_{3,5} & x_{4,5} & x_{5,5} & x_{6,5} & x_{7,5} \\
x_{1,6} & x_{2,6} & x_{3,6} & x_{4,6} & x_{5,6} & x_{6,6} & x_{7,6} \\
x_{1,7} & x_{2,7} & x_{3,7} & x_{4,7} & x_{5,7} & x_{6,7} & x_{7,7} \\
x_{1,8} & x_{2,8} & x_{3,8} & x_{4,8} & x_{5,8} & x_{6,8} & x_{7,8}
\end{array}
$$

Hamming weight and distance. The Hamming weight of a binary word, denoted by $w(.)$, is the number of nonzero elements in the word. Thus, if (1, 0, 1, 1, 1, 0) is such a word, then $w(1, 0, 1, 1, 1, 0) = 4$.

Interleaving essentially changes the channel from a burst-error channel to one with random errors by distributing the effects of an error burst over several data words.. Interleaving such as the example just presented is effective for protecting coded communications for the duration of the interleaved data word. It should be remembered, however, that interleaving does not decrease the long-term BER because the errors still occur [18].

The Hamming distance, *d*, is the minimum number of bit positions in encoded words by which any two code words are different. For example, if the code consists of the two words (1, 0, 1, 1, 1, 0) and (0, 0, 1, 1, 0), then $d = 3$ since the two words are different in the first and last two bit positions. The performance of a code is partially specified by d as follows:

1. The number of errors that can be detected is less than or equal to d.
2. The number of errors that can be corrected is less than or equal to $\lfloor (d-2)/2 \rfloor$ where $\lfloor x \rfloor$ denotes the greatest integer less than x.

Block codes. When there are k bits in a data word to be sent, and n bits are sent, then $n - k$ bits are added to each of the words. These $n - k$ bits are check bits selected in a particular way so that error control is maintained. There are 2^k possible messages to be sent and there are 2^n possible encoded words that are transmitted. The 2^k messages are called a block code and each of the possible 2^n transmitted symbols is called a code word.

While there are many kinds of codes available, the example included here is from the linear, systematic block codes. In these types of codes, parity bits are obtained from modulo-two addition of the data bits (linear), and the resultant parity bits are appended to the end of the data bits (systematic).

Hamming codes. As noted above, it is possible to organize data into a matrix and assign parity bits to the rows and columns of this matrix. Such an arrangement facilitates some error detection and error correction. This notion can be generalized where parity bits are applied to various combinations of the data bits (not just the rows and columns). If the number of data bits is given by d and the number of parity bits is given by p, then $d + p + 1 \leq 2^p$.

The Hamming code word, c, is the p bits appended to the d bits for each word. Such a code is identified as a (c, d) Hamming code. A Hamming code is identified by a generator matrix **G**, given by

$$\mathbf{G} = [\mathbf{I} : \mathbf{P}] \tag{6.46}$$

where **I** is the identity matrix and **P** specifies the particular Hamming code. The data words given by $d = [d_1, d_2, d_3, d_4]$ are multiplied by **G** to form the code words $\mathbf{c} = [d_1, d_2, d_3, d_4, p_1, p_2, p_3]$, that is, $\mathbf{c} = \mathbf{d}^{\mathrm{T}}\,\mathbf{G}$ (all math is modulo 2). The columns of **P** are selected so that each column is unique, thereby forcing the parity calculation to be performed over different sets of the data bits, and no resultant row is zero.

An example of a (7, 4) Hamming code is given by the generating matrix

$$\mathbf{G} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix} \tag{6.47}$$

and suppose that the following data word is to be encoded

$$\mathbf{d} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \end{bmatrix} \tag{6.48}$$

then the following code word is generated

$$\mathbf{c} = [1 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 0] \tag{6.49}$$

Clearly the first 4 bits are the original data word, while the last 3 bits form the appended parity bits.

At the receiver, the incoming data stream is checked by a matrix given by

$$\mathbf{H} = [\mathbf{P}^{\mathrm{T}} \quad \mathbf{I}] \tag{6.50}$$

to form a syndrome, s, for the data word, $\mathbf{s} = \mathbf{H}\ \mathbf{c}$. If $\mathbf{s} = 0$, then the data was received correctly, and if $\mathbf{s} \neq 0$, then an error has occurred.

In this example,

$$\mathbf{s} = \mathbf{Hc} = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{6.51}$$

(Remember that the arithmetic is carried out modulo two.) On the other hand, if [1001010] were received, then

$$\mathbf{s} = \mathbf{Hc} = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \qquad (6.52)$$

indicating an error has occurred in transmission.

Cyclic codes. A code is called cyclic when $(x_0, x_1, \ldots, x_n)$ is a code word whenever $(x_1, x_2, \ldots, x_n, x_0)$ is a code word. The latter is called a permutation of the former. BCH codes, developed by Bose, Chaudhuri, and Hocquenghem, are cyclic codes. Reed-Solomon codes are one form of BCH codes. They are used in CD players.

Convolutional codes. Convolutional code words are generated by combining sequences of bits together. Thus, the current code word depends on the current bit to be encoded as well as some number of previous bits. The bits are sequentially loaded into a shift register and the stages of the shift register are combined modulo-two in such a way to generate the convolutional code. The length of the shift register is called the constraint length, denoted as K, while the ratio of the number of bits in the input data word to the number of bits in the output, denoted by v, is called the rate. Such a convolutional encoder is shown in Figure 6.48. Here the constraint length $K = 3$ while the rate $= 1/v = 1/2$ since two output bits are generated for each input bit. Here

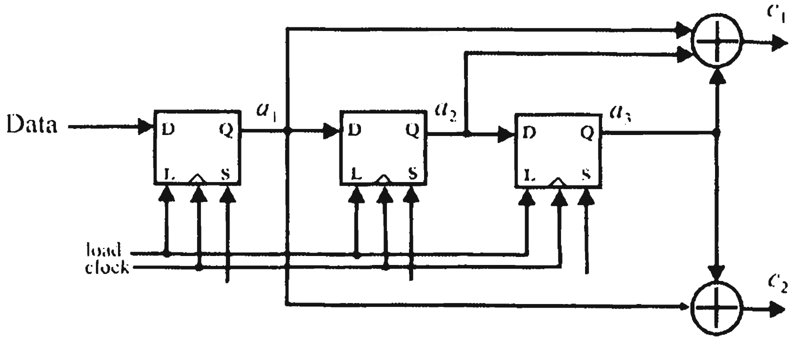


Figure 6.48 Convolutional encoder where $K = 3$ and rate $r = 1/2$.

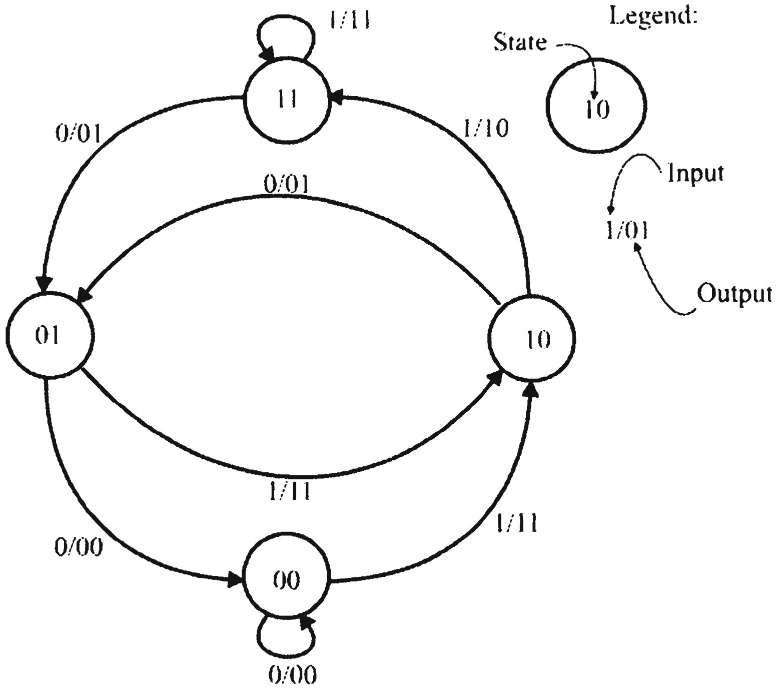


Figure 6.49 Finite state machine for the encoder shown in Figure 6.48.

$$c_1 = a_1 + a_2 + a_3 \qquad (6.53)$$

$$c_2 = a_1 + a_3 \qquad (6.54)$$

The initial seed for this shift register is zero.

Convolutional coders such as these are finite state machines, where the next state is determined by the input data and the current state. The state machine for this example is shown in Figure 6.49. The state of this machine is defined as (a_1, a_2) while the output is (c_1, c_2).

A trellis diagram is a way of displaying the operation of a convolutional encoder. The trellis diagram for the encoder shown in Figure 6.48 is shown in Figure 6.50. In this diagram, *x*/*yy* refers to *x* = input and *yy* = output after the next clock pulse loads the input into the first stage and all the other stages are shifted to the right one place.

Viterbi Decoding. Viterbi decoding is a form of maximum-likelihood decoding. It is implemented with a convolutional decoder. The concept of maximum-likelihood decoding arises because of the Hamming distance notion mentioned above. If a noncode word is received, the most likely code word that was sent is the one with the minimum Hamming distance from the received word. This is perhaps best illustrated

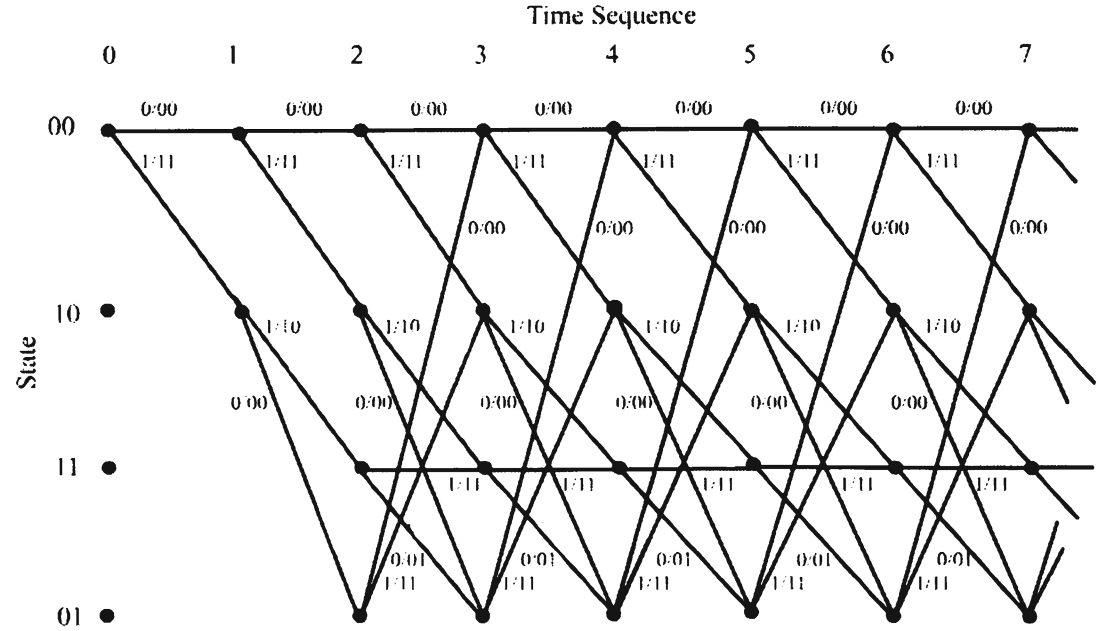


Figure 6.50 Trellis diagram for the encoder in Figure 6.48.

with an example. Suppose that the code under consideration is of length three. The probability there are no errors is given by probability bit 1 is error-free × probability bit 2 is error-free × probability bit 3 is error-free $= p \times p \times p = p^3$ since by the binary symmetric channel assumption bit errors are independent of each other. Likewise, there are three ways of having 1 bit in error and the probability of this is given by

$$\Pr\{1\text{bit in error}\} = \frac{\begin{array}{l}\Pr\{\text{bit 1 wrong and bits 2 and 3 error free}\}\\ \quad +\Pr\{\text{bit 2 wrong and bits 1 and 3 error free}\}\\ \quad +\Pr\{\text{bit 3 wrong and bits 1 and 2 error free}\}\end{array}}{3} \tag{6.55}$$

$$= \frac{(1-p)\times p\times p+(1-p)\times p\times p+(1-p)\times p\times p}{3} \tag{6.56}$$

$$=(1-p)p^2 \tag{6.57}$$

Lastly, the probability of there being 2 bits in error is given by

$$\Pr\{2\,\text{bits in error}\} = \frac{\begin{array}{l}\Pr\{\text{bit 1 and bit 2 wrong and bit 3 error free}\}\\ \quad +\Pr\{\text{bit 1 and bit 3 wrong and bit 2 error free}\}\\ \quad +\Pr\{\text{bit 1 error free and bits 2 and 3 wrong}\}\end{array}}{3} \tag{6.58}$$

$$= \frac{(1-p)\times(1-p)\times p+(1-p)\times(1-p)\times p+(1-p)\times(1-p)\times p}{3} \tag{6.59}$$

$$= (1-p)^2 p \tag{6.60}$$

Suppose $p = 0.9$. Then P(no errors) = 0.73, P(1 bit in error) = 0.08 and P(2 bits in error) = 0.009. Therefore, the maximum probability, if there are errors present, corresponds to 1 bit in error—that is, the code word that is a distance of one away from the received word. Thus, the most likely code word is the one with minimum Hamming distance from the code word received.

If the code words are short, it is common practice to compare the received coded words, possibly with errors, to a lookup table to find the word that is the closest to the received word. For long code words such a lookup technique is impractical. The Viterbi decoding algorithm was devised as a more efficient way of doing this comparison. What the Viterbi decoding process does is prune the list over which the search is performed upon receipt of a code word. The most likely code words are kept for a period of time. When a new coded symbol is received, the Hamming distance between the sequence made up of the concatenation of the new symbol with the received symbols and each surviving path through the trellis is calculated. Those sequences with the smallest Hamming distance are kept for the next symbol period, and the process is repeated. After a certain specified number of symbol periods, a decision is made about the received sequence of symbols. The operation of the algorithm is best understood by considering an example. Suppose in the above encoder the depth of code words at the receiver is set at five symbols and the maximum Hamming distance is set at $d = 4$ bits. Then through the first five symbols received, all of the possible sequences are kept. These are shown in Table 6.7 for the example received sequence shown. When the sixth symbol is received, the first symbol is dropped and only those symbols with a Hamming distance less than five are kept. These are shown in Table 6.8. Note that the Hamming distance is the total distance, even though the symbols received older than five are dropped. The total distance is easily calculated since the difference in the last symbol received and the possible symbols from the trellis is simply added to the running total.

In this example, the maximum Hamming distance was set at $d = 4$ in order to have the sequence retained. It may make more sense to adopt the strategy of keeping

Table 6.7 First Five Time Intervals for the Decoder Corresponding to the Encoder Shown in Figure 6.42

Time interval	00	1	2	3	4	5	6	7	
Transmit data		1	0	1	1	0	1	0	
Encoder state	00	10	01	10	11	01	10	01	
Error-free coded sequence		11	01	11	01	01	11	01	
Received sequence		11	01	10	01	11			
Surviving code words							Previous state	Current state	Hamming distance
		00	00	00	00	00	00	00	7
		00	00	00	00	11	00	10	5
		00	00	00	11	00	10	01	7
		00	00	00	11	10	10	11	6
		00	00	11	00	00	01	00	7
		00	00	11	00	11	01	10	5
		00	00	11	10	01	11	01	7
		00	00	11	10	11	11	11	6
		00	11	10	01	00	01	00	5
		00	11	10	01	11	01	10	3
		00	11	10	11	01	11	01	5
		00	11	10	11	11	11	11	4
		00	11	00	00	00	00	00	7
		00	11	00	00	11	00	10	5
		00	11	00	11	00	10	01	7
		00	11	00	11	10	10	11	6
		11	00	00	00	00	00	00	5
		11	00	00	00	11	00	10	3
		11	00	00	11	00	10	01	5
		11	00	00	11	10	10	11	4
		11	00	11	00	00	01	00	5
		11	00	11	00	11	01	10	3
		11	00	11	10	01	11	01	5
		11	00	11	10	11	11	11	4
		11	10	01	00	00	01	00	7
		11	10	01	00	11	01	10	5
		11	10	01	11	00	10	01	7
		11	01	01	11	10	10	11	4
		11	10	11	01	00	01	00	5
		11	10	11	01	11	01	10	3
		11	10	11	11	01	11	01	5
		11	10	11	11	11	11	11	4

Table 6.8 Sixth Time Interval for the Encoder Shown in Figure 6.42

Time interval		1	2	3	4	5	6	7		
Transmit data			0	1	1	0	1	0		
Encoder state	00	10	01	10	11	01	10	01		
Error-free coded sequence			01	11	01	01	11	01		
Received sequence			01	10	01	11	11			
								Previous state	Current state	Hamming distance
Surviving code words			11	10	01	11	00	10	01	5
			11	10	01	11	10	10	11	4
			11	10	11	11	01	11	01	5
			11	10	11	11	11	11	11	4
			00	00	00	11	00	10	01	5
			00	00	00	11	10	10	11	4
			00	00	11	10	01	11	01	5
			00	00	11	10	11	11	11	4
			00	11	00	11	00	10	01	5
			00	11	00	11	10	10	11	4
			00	11	10	11	01	11	01	5
			00	11	10	11	11	11	11	4
			01	01	11	10	01	11	01	5
			01	01	11	10	11	11	11	4
			10	11	01	11	00	10	01	5
			10	11	01	11	10	10	11	4
			10	11	11	11	01	11	01	5
			10	11	11	11	11	11	11	4

the sequences with the three smallest Hamming distances, for example, than to set d at a fixed number.

Per-survivor processing. A recent addition to the field of maximum-likelihood decoding convolutional codes is called per-survivor processing (PSP) [19]. Depending on the decoding scheme used, most decoders require as inputs some or all of the parameters associated with a received signal. These parameters could include the SNR and the carrier phase. Therefore, they are known, assumed, or estimated in a global sense and then input to the decoding algorithm, such as the aforementioned Viterbi algorithm, for all the paths through the trellis.

In PSP this is not the case. The channel is assumed to be unknown. In fact, the channel can be time varying. Each surviving path through the trellis retains its own estimate of these parameters and they are updated with each new symbol received. It is claimed that significantly better performance ensues from PSP than other conventional types of decoders.

Soft decoding. In the decoders described so far, at some point a decision is declared that a decoded bit is a one or a zero or that a specific sequence has been received. This is called hard decision decoding. In maximum-likelihood decoding, however, likelihood ratios are computed based on the received bits. These ratios are probabilities with values between zero and one. If these values are output from the decoder and used to make decisions about the received bit as opposed to having the decoder output a one or zero, it is called soft decision decoding.

In hard decoding some of the information available within the decoding process is not made available to the decision processes that decide which bit was received. This information loss results in degraded performance. Soft decision decoding typically improves error performance by about 2 dB. That is, the same BER performance is obtained with 2 dB less SNR.

Concatenated codes. These coding techniques use one code followed by another, called outer and inner coders/decoders. Interleaving is also usually employed to increase the bit error performance. The error detecting and correcting effects are different.

Turbo codes. One of the classic problems with employing coding techniques has been the complexity of the decoder. Code generators are normally simple devices, but as codes get more complex, the decoders become complicated as well.

Turbo codes do not have this difficulty. They use what is referred to as maximum a posteriori (MAP) processing. If $\Pr(A)$ is the a priori probability of event A occurring, an event in this case being the transmission of code word A, that is $\Pr(A)$ is the

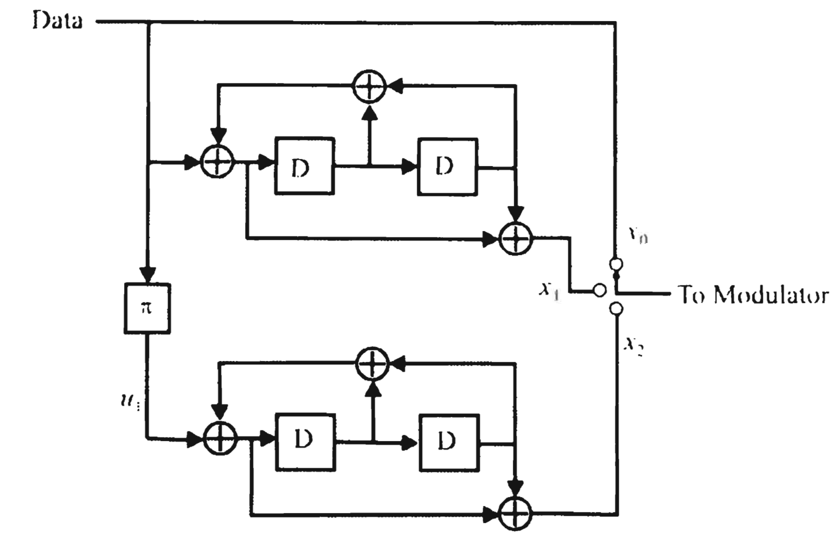


Figure 6.51 An example of a Turbo encoder.

probability of A occurring at all, then Bayes' theorem says

$$\Pr(A|B) = \Pr(B|A)\Pr(A)/\Pr(B) \tag{6.61}$$

$\Pr(A|B)$ is the a posteriori probability of event A occurring given that event B was observed. $\Pr(B|A)$ is called the likelihood function and Pr(B) is called the evidence. The MAP criteria selects the A for which the a posteriori probability is maximum.

These codes are also called parallel-concatenated convolution codes. The encoder is made up of two or more convolutional encoders, each of which implements a constituent code. Interleavers, denoted here by π, interconnect the constituent encoders. The interleavers approximate what is known as a uniform interleaver that perfectly scrambles the input sequence. The code words generated by the encoder consist of the original data input to the first encoder followed by the bits from the constituent encoders, in the appropriate order. Normally the systematic bits from any but the first constituent encoder are not transmitted. An example of a Turbo encoder is shown in Figure 6.51 [20].

The decoder is comprised of one constituent decoder module for each encoder module. The decoders perform soft decision decoding of their input sequences, and the same inter-leavers used in the encoder interconnect them. It has been reported that BERs of 10^{-6} with $E_b/N_0 = -0.06$ with code rates of 1/15 have been achieved with Turbo code implementations through simulations [21].

Reed Solomon Codes. Reed Solomon (RS) codes are systematic linear block codes, part of a subset of the BCH codes called *nonbinary BCH.* They are block codes

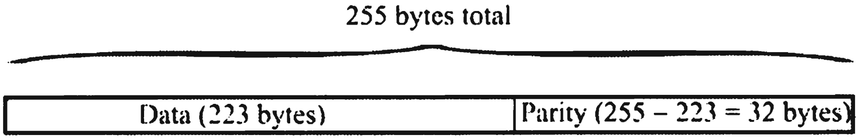


Figure 6.52 Reed Solomon example RS(255, 223)

because the original message is split into fixed length blocks and each block is split into *m* bit symbols. They are also linear because each *m* bit symbol is a valid symbol and since each code word contains the original daa they are systmatic codes. The transmitted codeword contains the original data with extra CRC or "parity" bits appended. These codes are specified as RS(*n*, *k*), with *m* bit symbols. This means that the encoder takes *k* data symbols of *m* bits each, appends *n* – *k* parity symbols, and produces a code word of *n* symbols (each of *m* bits). An example of this coding is illustrated in Figure 6.52 where RS(255, 223) is shown.

Reed Solomon codes are based on Galois fields (GF, fintie fields) [22]. These fields are of the form GF(p^m), where *p* is any prime. RS makes use of Galois fields of the form GF(2^m), where elements of the field can be represented by *m* binary bits. Hence, RS codes of the form RS (2^8) lend themselves well to digital communication.

Several error-correcting techniques are compared to the Shannon limit in Figure 6.53 [23]. The Shannon capacity limit capacity of a communications channel is the theoretical maximum information transfer rate of the channel, for a particular noise level given by (6.22). White Gaussian noise is assumed for this data. The turbocoded channel (TPC) is seen to exceed the performance of all the other coding techniques considered and is only about 1 dB or so away from the Shannon limit for this

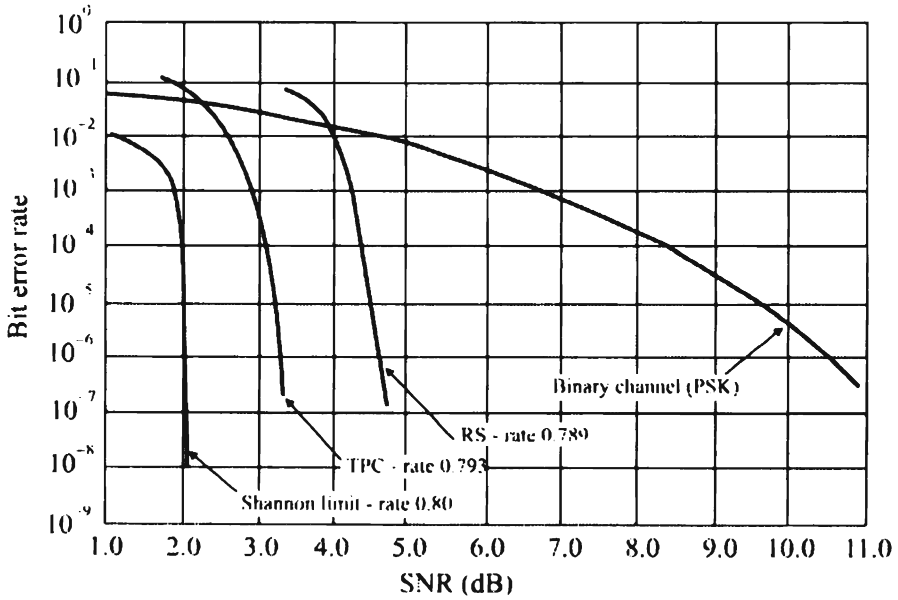


Figure 6.53 Comparison of several coding techniques with the Shannon limit. (*Source:* [21], ©EE Times 1998. Reprinted with permission.)

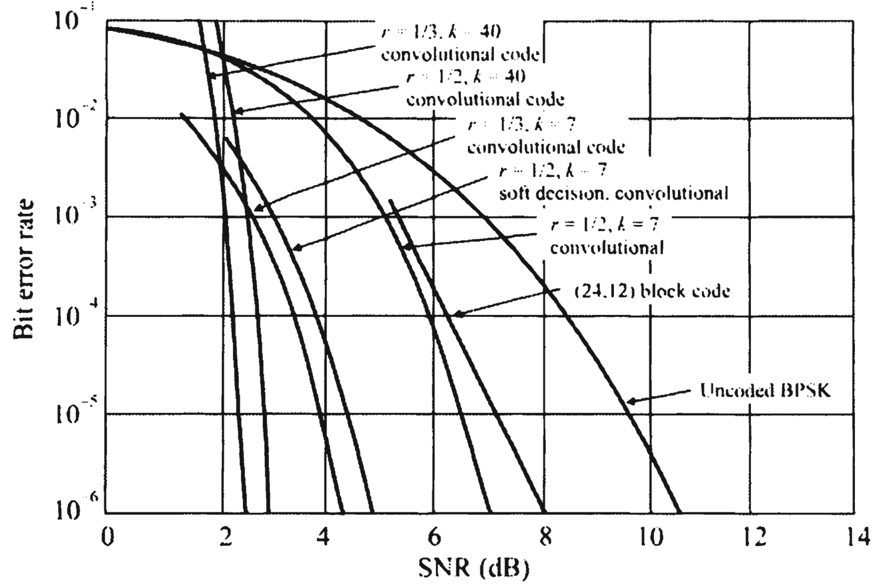


Figure 6.54 BER performance of some codes with BPSK. The BER drops off faster as the SNR is increased if the data is encoded. (*Source:* [22], © John Wiley & Sons 1988. Reprinted by permission.)

example. It is about 1.5 dB better in performance for most BER < 10^{-4} as compared to the next best code, the Reed Solomon (RS) code. The uncoded channel case is also shown as the binary channel (PSK).

The advantages of coding as described above can be seen in Figure 6.53 [24] where the probability of a bit error, or equivalently, the BER, is plotted versus the SNR as specified by the E_b/N_0 for several coding schemes. As shown, very good performance is possible for low SNRs over uncoded data streams.

BCH Coding

For all positive integers, $m \geq 3$ and $p < 2^{m-1}$ there is a binary BCH code [25], with the appellation (n k, p) BCH code, with the following parameters:

- Block length: $n = 2^m - 1$;
- Number of parity check bits: $n - k \leq mp$; and
- Minimum distance: $d_{\min} \geq 2p+1$.

Each such code can correct up to p-bit errors, and therefore it is also referred to as a p-error-correcting code. The error performance of a few popular BCH codes are illustrated in Figure 6.54, compared with no error coding.

6.8 Modems

A modem, a term that is formed from the concatenation of modulator and demodulator, is a device that converts information from the digital domain into the analog domain and back again. Data must be in analog form if it is to be transmitted over much of the existing worldwide telecommunications infrastructure.

Modems are primarily used to communicate between digital devices—principally computers—over analog networks. They are key components of wide area networks, but the data rate they can transmit is relatively slow. Although initially used for wireline communications, they can also be used with radios in wireless networks. Since wireless environments are typically much more error-prone, some form of error detection and correction is necessary. Historically these capabilities have not normally been part of the modem, but are inserted into the data stream before the stream is sent to the modem for transmission. With convolutional coding as represented by Trellis code modulation and Viterbi decoding, that picture is changing.

Standards, or specifications, have been developed so that interoperability between modems from different manufacturers can be accomplished. The International Telecommunication Union-Telecommunications Standardization Sector (ITU-TSS), or ITU-T for short, is an agency of the United Nations that has traditionally been the governing body for such standards, by mutual agreement. Such specifications are called recommendations by the ITU. Table 6.9 lists the recommendations that apply to modem communications. All but the last recommendation transmit digital data with analog techniques over modems. V.90, on the other hand, does not convert the digital data to analog forms. The digital data is sent over the channel in digital form.

Modems built according to the higher data rate specifications are backward-compatible with the lower data rates, and frequently modems can handle several other specifications as well. These specifications provide for training for the channel, which is a form of equalization. Training sequences are used for this, where the data being sent is known to the receiving modem. The received sequence is examined for accuracy and the ISI is determined. If there is too much interference or other degradation, such as noise, the sending modem is notified to lower the data rate. In this way, handshakes between the sender and receiver establish the highest rate that the channel can support and that is the rate used for the message transfer.

According to Shannon's theorem given earlier (6.1), the maximum capacity of a memoryless voice grade channel corrupted by only Gaussian noise of (guaranteed) 2,700 Hz bandwidth is given in Table 6.10 for various SNRs. The phone system in the United States has a specified (theoretical) noise-limited dynamic range of 39.5 dB, but more typical values are in the 30–35 dB range. Since the guaranteed bandwidth is only 2,700 Hz, data rates should be expected to be 27 Kbps or less. SNRs are typically substantially less than this for RF links, especially in the VHF range.

Table 6.9 Characteristics of Many of the ITU Modem Recommendations

Recommendation	Bit Rate (bps)	Baud Rate (sps)	Modulation	Coding
V.17	7,200	1,200	QAM	None
V.17	9,600	1,200	QAM	None
V.17	12,000	1,200	QAM	None
V.17	14,400	1,200	QAM	None
V.21	300	300	FSK	None
V.22	1,200	600	PSK	None
V.22bis	1,200	600	QAM	None
V.22bis	2,400	600	2^4QAM	None
V.23	600	600	FSK	None
V23	1,200	1,200	FSK	None
V26	1,200	1,200	BPSK	None
V26	2,400	1,200	2^2PSK	None
V.27bis	2,400	1,200	2^2PSK	None
V.27bis	4,800	1,200	2^3PSK	None
V.27ter	2,400	1,200	2^2PSK	None
V.27ter	4,800	1,600	2^3PSK	None
V.29	4,800	2,400	2^2PSK	None
V.29	7,200	2,400	2^3QAM	None
V.29	9,600	2,400	2^4QAM	None
V32	2,400	2,400	BPSK	None
V.32	4,800	2,400	QPSK	None
V.32	9,600	2,400	2^4QAM	None
V.32	9,600	2,400	2^5QAM	Trellis (2 dim)
V.32bis	4,800	2,400	QAM	Trellis (2 dim)
V.32bis	7,200	2,400	2^5QAM	Trellis (2 dim)
V.32bis	9,600	2,400	QAM	Trellis (2 dim)
V.32bis	12,000	2,400	QAM	Trellis (2 dim)
V.32bis	14,400	2,400	2^6QAM	Trellis (2 dim)
V.32terbo	16,800	2,400	QAM	Trellis (2 dim)
V.32terbo	19,200	2,400	2^9QAM	Trellis (2 dim)
V.33	14,400	2,400	2^7QAM	Trellis (2 dim)
V.34	28,800	2,400–3,249	2^8–2^9QAM	Trellis (4 dim)
V.34bis	33,600	2,400–3,249	2^8–2^9QAM	Trellis (4 dim)
V.90	56,000	N/A	PAM	None

sps = samples per second, bps = bits per second

Table 6.10 Channel Capacity of a 2,700-Hz Channel

SNR (dB)	Channel Capacity (bps)
5	5,590
10	9,340
20	17,977
30	26,912
40	35,877

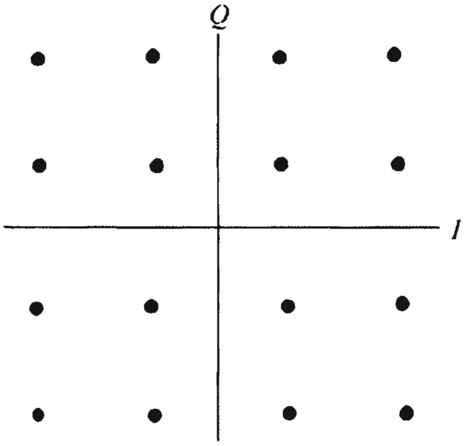


Figure 6.55 Constellation for V.32.

Because of this, frequently coding schemes are employed in modems to increase the data rate. Trellis coding (TCM) is a popular choice. Using TCM, modems have been able to achieve 8 bits per symbol to provide a data rate of 19.2 Kbps at a channel signaling rate of 2,400 symbols per second. This is four times better than GMSK or π/4 QPSK.

Unencoded QAM at 9,600 bps over a 2,400-baud voice grade channel yields a 14 dB advantage as compared to TCM at the same rate. TCM signals have about the same BER performance at 19,200 bps over this channel.

V.32 at 9,600 bps specifies a convolution code for its TCM which facilitates full-duplex communication over a two-wire circuit. A coding gain of 4 dB is achieved in V.32. Details on the V.32 are presented here to illustrate a typical encoding scheme. The two principal modes in V.32 facilitate either 4,800 bps or 9,600 bps data rates, both at a 2,400 full-duplex baud rate over two wires. The former is for unencoded transmission, while the latter implements TCM. The 16QAM constellation for the 4,800-bps rate is shown in Figure 6.55, while in Figure 6.56 the 2^5QAM constellation for 9,600 bps is shown.

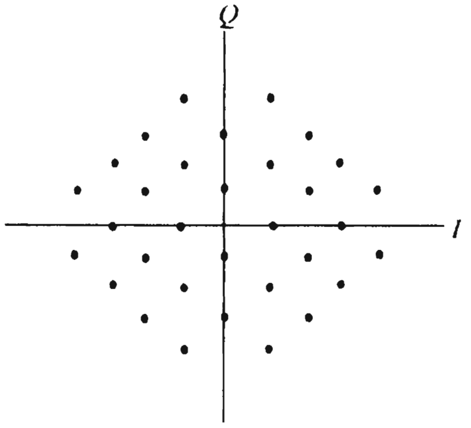


Figure 6.56 Constellation for V.36.

For 2^4QAM, 4 data bits are combined into each one of the constellation points for transmission. For 2^5QAM, obviously 5 bits are needed, but these 5 bits are constructed from the same 4 data bits. The fifth bit is a result of the TCM added to the data stream for error control. The first 2 data bits, denoted Q1 and Q2, are first differentially encoded via the formulas

$$Y1_n = Q1_n \oplus Y1_{n-1} \tag{6.62}$$

$$Y2_n = (Q1_n \oplus Y1_{n-1}) \oplus Y2_{n-1} \oplus Q2_n \tag{6.63}$$

(n here refers to the bit interval number).

This encoding guards against a 180° phase ambiguity in the channel. The last two data bits, $Q3$ and $Q4$, are not encoded. $Y1$ and $Y2$ are then used to form $Y0$, which carries the FEC information. $Y0$ is the product of a $r = 2/3$ $K = 3$ convolutional encoder.

The combinational logic in this encoder, as well as other TCM encoders, is designed according to rules defined by set composition on the constellation for decoding. The Euclidean distance between constellation members in any given set is maximized at each step of the decomposition process.

At the demodulator a Viterbi decoding algorithm is used with V.32 as mentioned earlier. This is a convolutional code decoder that minimizes the mean squared error.

In V.34, data rates higher than 28,800 bps are possible. Rates of 33,600 bps and 38,400 bps are theoretically possible, but ultra clean and balanced lines are necessary. When V.42 is combined with V.34, rates up to 115,000 bps are theoretically possible. Normal telephone lines have a bandwidth that is nominally 300–3,000 Hz. Small changes in these frequencies can have significantly deleterious effects on the data rate possible over these lines. The necessary bandwidths for transmitting these high data rates are given in Table 6.11. Modern modems can adjust these bandwidths to be compatible with the actual PSTN lines being used by exchanging training sequences that have a known bit pattern to be measured as well as by probing the channel. Probing is transmitting and measuring the response of the channel at known frequencies and levels. ISDN is expected to change that picture dramatically since no modems are required—the communication is digital in origin and remains that way throughout the path. Interface devices will be required, however. While the modem recommendations discussed above are not inextricably connected to the Internet, the V.90 recommendation was inspired by it. A modem modulates and demodulates low-frequency carriers with digital data using some form of QAM. This inherently limits their speed to less than about 30 Kbps due to quantization noise in the A/D conversions involved. The downstream leg of the V.90 is not modulated this way. It is sent in digital form, as if the PSTN were a digital network. The information is pulse

Table 6.11 Bit Rates Possible Through Band-Limited PSTN Channels

Symbol Rate	Carrier Frequency	Bandwidth Requirements	Maximum Bit Rate
2,400 sps	1,600 Hz	400–2,800 Hz	21,600 bps
	1,800 Hz	600–3,000 Hz	21,600 bps
2,743 sps	1,646 Hz	274–3,018 Hz	24,000 bps
	1,829 Hz	457–3,200 Hz	24,000 bps
2,800 sps	1,680 Hz	280–3,080 Hz	24,000 bps
	1,867 Hz	467–3,267 Hz	24,000 bps
3,000 sps	1,800 Hz	300–3,300 Hz	26,400 bps
		500–3,500 Hz	26,400 bps
		375–3,376 Hz	26,400 bps
3,200 sps	1,829 Hz	229–3,429 Hz	28,800 bps
	1,920 Hz	320–3,520 Hz	28,800 bps
3,429 sps	1,959 Hz	244–3,674 Hz	28,800 bps

sps = samples per second, bps = bits per second

amplitude modulated (PAM) onto the line in the downstream direction, from the service provider to the customer's premises. In this way, depending on the quality of the phone line, rates up to 56 Kbps are possible. Since the majority of high-speed traffic on the Internet is downstream, with the upstream normally consisting of slow keystrokes, for example, substantially higher throughput over the Internet is possible with V.90. The analog upstream leg in V.90 is V.36.

In fact, the vast majority of the PSTN in the United States is digital. In most populated places the only remaining analog portions are on the local loop. This is the part of the network that connects the local office with the residences and businesses. The remainder of the PSTN is digital. This statement does not necessarily apply in countries other than the United States, however.

6.9 Facsimile

Facsimile (fax) communication is the transfer of images—typically typed pages—over analog communication systems. Group I fax is an obsolete standard where either AM or FM was used as the modulation scheme. In the former, the blacker the pixel, the higher the amplitude of the carrier; while in the latter, the blacker the pixel, the higher the modulating tone. Group II fax is also an old standard that is almost obsolete. It uses vestigial sideband AM as its modulation scheme, and the whiter the pixel, the higher the tone that was sent. These standards have been almost totally replaced by Group III fax.

Run-length encoding is used in the Group III fax protocol. It takes each line of pixels and run-length encodes them before transmission. This is one-dimensional encoding in that each line is encoded, as opposed to the whole page. Two-dimensional encoding encodes the first line with run-length encoding as just described, but subsequent lines are encoded as differences from the previous line. Significant compression can result since there is frequently quite a bit of correlation between lines of text. The specification is error-detecting but not error-correcting; therefore, errors can propagate down the page this way. Because of that, a limit of two lines is encoded this way for standard-resolution faxes and four for high-resolution faxes.

The Group IV fax standard covers facsimile transmissions over ISDN at 64 kbps. In this case the compression is the same as just described for two-dimensional compression except that there is no limit on the number of lines that are encoded based on the first line.

Within each of these groups there are classes defined, which essentially describes the amount of processing that a modem must do to transmit and receive facsimile messages. In class 1 most of the work must be done by the computer, and the fax modem work is limited to modulation of the carrier and asynchronous-to-synchronous conversions. In the class 2 standard, the modem is smarter and must do more of the processing. In particular the modem is responsible for modulation and asynchronous-to-synchronous conversion, as well as taking care of the fax protocol. The computer is responsible only for the compression of the image (as well as overall management, of course).

6.10 Communication Security

One of the most significant issues associated with information transport is the security, or privacy, of the information. This has been traditionally true of military communications but is increasingly becoming important in the private sector as well. Current, unclassified techniques to deal with this problem are discussed in this section.

6.10.1 Data Encryption

Encryption is the most common way to protect information that is to be transported from one place to another. Information is encrypted when a mathematical operation is performed on it to make it unreadable by unauthorized persons. Cryptography is an electronic protect measure for the benefit of whoever uses it. It refers to the process of transforming information from its current state to another where the content is hidden, and from which, with the proper knowledge of such components as encryption keys, the original content can be retrieved and without which the data cannot be retrieved. Any information in any state can be encrypted. The most well known perhaps is the encryption of written or radio communication. Other types include encrypting data on a disk of a PC and parents verbally spelling words that their 2-year-old child cannot understand. Sir Arthur Conan Doyle endowed Sherlock Holmes with a special skill in dealing with cryptography.

Historically, this is an area that has been dominated by the government, but recently the private sector has become more interested. Techniques have been devised to protect information, while making it relatively easy to move the encryption keys among legitimate communication nodes.

Authentication is the process of verifying the identity of the generator of some data and, therefore, the integrity of that data. A principal is the party whose identity has been identified. A verifier is the person who wants the verification. Data integrity means that there is assurance that the data received is the same as the data generated. Confidentiality refers to protection from disclosure of information to those unauthorized to receive it. Authorization is the process used to ensure that a principal may perform an operation.

User nonrepudiation refers to the inability of the source of the data to deny that the data was generated by that source. That is, if the process is correct and it identifies a particular source of the data, then certainly that is the source of the data.

6.10.2 Public Key Encryption

In public key encryption, each entity i has a public key U_i, known to everyone, and a private key R_i, known only to the entity. A message to i is encrypted by using U_i while R_i is used by i to decrypt the message and, if encrypted with a user's public key, it can only be decrypted with the user's private key.

Public key encryption is asymmetric encryption since each end of a link uses a different cryptologic variable to encrypt and decrypt data. This is as opposed to symmetric encryption where the same cryptologic variable is used at both ends. The excerpt in Figure 6.57, taken from [25], describes by example how this encryption technique works.

Because of the extensive (and therefore slow) computations involved in using public key cryptography, most of the time it is used to exchange session cryptologic keys so that communications can proceed in a symmetric manner, which is much faster. It is also used for exchanging digital signatures.

Session keys are the equivalent to what is called the one-time-pad. Such a scheme uses a crypto key only once (for the current session). After that, it is discarded. Such techniques are virtually impossible to decrypt since the keys are only used once.

6.10.3 Digital Signatures

A security issue that has arisen because of the mass movement toward electronic commerce (for example, retail sales over the Internet) is how one knows that the

Encryption 101

Public-key cryptography uses different keys for encrypting and decrypting, which are created as shown below, and then assigned to each user. Whenever someone wants to send a message, he looks up the recipient's public key and uses it to encrypt the message. The recipient then uses his private secret key to decrypt the message. For simplicity's sake the example below uses small prime numbers; in practice, the keys are very large numbers, making it difficult for outsiders to factor N and decipher the message. Both sender and receiver must have the same encrypting algorithm in hardware or software in their computer. In this example, adapted from the US Office of Technology Assessment's 1987 report Defending Secrets, Sharing Data, the algorithm is based on the public-key RSA algorithm, named after its three Massachusetts Institute of Technology inventors—Ronald Rivest, Adi Shamir, and Leonard Adelman.

Creating the public key	
1. Pick an odd number, E.	E = 5
2. Pick two prime numbers, P and Q, where (P-1)(Q-1)-1 is evenly divisible by E.	P = 7, Q = 17
3. Multiply P and Q to get N.	N = P x Q = 7 x 17 = 119
6. Concatenate N and E to get the encrypting or public key.	Public Key = NE = 1195
Creating the private key	
1. Subtract 1 from P, Q, and E, multiply the results, and add 1.	(P-1)(Q-1)(E-1)+1 = 6 x 16 x 4 + 1 = 385
2. Divide result by E to get D.	D = 385/5 = 77
3. Concatenate N and D to get the decrypting or private key.	Private Key = ND = 11977
Encrypting the message with the public key	
1. The message is converted to numerical equivalents. The letter S, for example, may be represented by 19.	Plain Text = 19
2. The algorithm:	
a) raise plain text to power of E.	19^5 = 2476099
b) divide by N.	2476099/119 = 20807 with a remainder of 66
3. The remainder is the encrypted value or cipher text.	Cipher Text = 66
Decrypting the cipher text with the private key	
1. The algorithm:	
a) raise cipher text to power of D.	66^{77} = 1.27 ... E140
b) Divide by N.	1.27 ... E140 / 119 = 1.069 ... E138 with a remainder of 19
2. The remainder is the decrypted value or plain text.	Plain Text = 19

Figure 6.57 Example of RSA public key encryption algorithm. (*Source:* [24], © 1989, IEEE. Reprinted with permission.)

person at the other end of a transaction is who he or she claims to be. In the old analog world of plastic credit cards and bank checks, it was a matter of checking the signature on the card or check and the picture on an identification. Techniques such as these are not as yet practical for use over the Internet, yet if electronic commerce is to survive, they will be.

Public key cryptology can be used to facilitate such checking, however. The approach takes the appellation digital signatures. In order to understand how they work, it is necessary to understand the functioning of hashing, which is a process of taking one number (A) and converting it into another (B), where the inverse (knowing B, find A) is next to impossible. The two most popular hashing algorithms are ones invented at MIT by Ron Rivest, called MD5, and one invented by NIST and NSA called the secure hash algorithm (SHA). The former generates a 128-bit number and the latter a 160-bit number. Of course, the larger the number of bits, the more secure the data.

A message that is to get through without being changed—or if it is changed, the change is detected at the receiving end—can be sent by hashing the contents of the message, encrypting this hashed message with a public key, and sending the encrypted hashed value along with the message. The receiver decrypts the hashed value with the sender's private key, hashes the message with the same algorithm that the sender used, and compares the hash value sent with the one calculated at the receiver's location. If they match, then the receiver is assured that the message arrived unchanged, while if they do not match, it can be virtually assured that the message did not come from the sender because of the uniqueness of the hashed value. The hash value is the digital signature since it is the only number that can be generated from the document; and furthermore, the encryption/decryption process only works accurately if the public and private keys are correct—if they are not, the hash value will decrypt incorrectly and the message will be garbled.

6.10.4 Data Encryption Standard

One of the ways to attack an encryption system is by brute force. Encryption algorithms are not normally kept secret and the reason for this is simple. If it is necessary to keep the algorithm secret then the algorithm itself is vulnerable—there is something about the algorithm that must be kept from would-be interlopers. Such weaknesses are characteristic of poor encryption systems. It is more normal to make the encryption key secret, but not the algorithm that uses the key. Brute force attacks on an encryption system are conducted using computing power, and every possible key combination is tried until one is found that yields the original message, or at least a message that makes sense.

We might ask how it is known if the original message is obtained or some other message, unless one already knows the original message. While this is an interesting

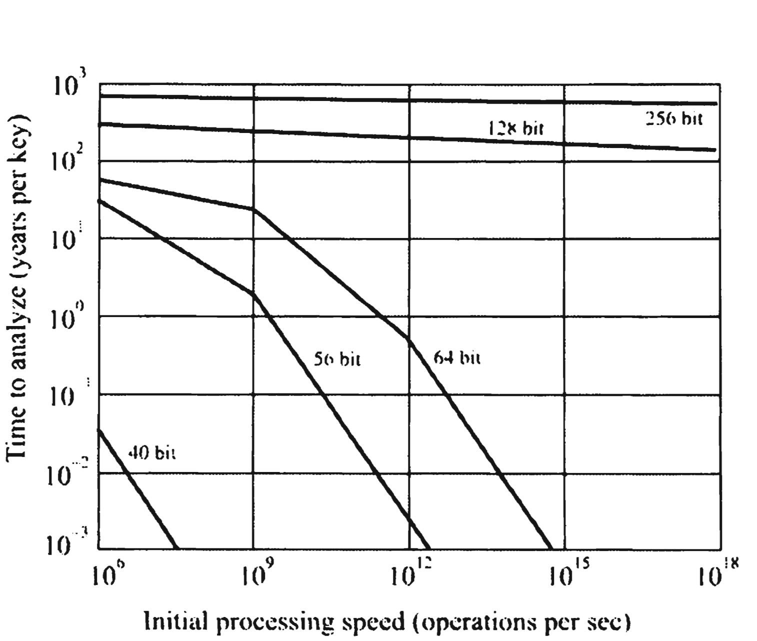


Figure 6.58 Time required to recover encryption keys. (*Source:* [26], © Artech House. Reprinted with permission).

paradox, it is normally the case that anything but the original message will be garbled and senseless, so it is obvious that one has obtained the original message.

The longer a key is, the safer the encryption is against a brute force attack. To see why this is true, if only eight bits are used for encryption, then 2^8 = 256 different values of the encryption key would have to be tried in order to find the correct key. If one value took 1 ms to try, then it would only take 256 ms (half this value on average) to break the code. On the other hand, if a key is 100 bits long, then $2^{100}/2$ different values would have to be tried. On the same computer, taking 1-ms per try, it would take on the order of 10^{19} years to break the cipher. The times required to attack encryption systems including the improvements over time due to Moore's law are illustrated in Figure 6.57 [25].

Parallel processing can be implemented that substantially reduces the equivalent of the 1-ms per try, however. Also, the 1 ms used in this example is much slower than what high-performance computers can do. Using realistic times and large numbers of parallel computers, cipher systems based on 100 bits can probably be broken by brute force in reasonable times, if the end result is judged to be worth the expense.

The data encryption standard (DES) was adopted in 1977 as the standard technique in the U.S. government for encrypting sensitive but unclassified information. [The U.S. Department of Defense (DoD), for example, does not use DES for classified data. It was intended to keep an individual's information private, including corporations.] It was made available to the private sector as a means for providing a reasonably secure method for protecting sensitive information. It uses a 56-bit key, which, at the time of its adoption, was a considerable amount of protection against the then-existing and projected computer speeds. This provided 72,057,594,037,927,936 possible different keys.

DES uses a block encryption technique as opposed to a stream cipher. This means that blocks of data 64 bits in size are encrypted at the same time. In a stream cipher 1 bit at a time is enciphered. Confusion and diffusion are the techniques at the heart of the algorithm, and the data is operated on 16 times by these two, which is determined by the key. The notion of diffusion is to spread the plaintext over the ciphertext, so that the inherent redundancy in the plaintext (assumed to be there, and if the plaintext is English text, for example, there is considerable redundancy) is hidden. A simple example of this is transposition of the characters in the plaintext. Confusion is the process of disestablishing the relationship of the plaintext from the ciphertext. A simple way to do this is by substitution. It is said that this technique was used by Julius Caesar, by shifting the alphabet by three places and substituting the characters accordingly. As mentioned, at each round through the encipherment process, confusions and diffusions are performed.

Another important part of the algorithm is permutations of the data, so that the avalanche criterion is met. This ensures that 1 bit of data affects at least two substitutions. In this way, if the ciphertext is changed in only one place, the decrypted plaintext is dramatically different from the pre-encrypted plaintext. The actual security of the algorithm is accomplished at a nonlinear processing stage.

The way such keys are used is typically to compute the exclusive OR of the data bits with the bits in the encryption key. In fact, the processing is more complex than this. A series of shifts and permutations are used that, being known to both the sender and legitimate recipient, can be undone properly. Since the key is unknown to a third party, the resultant data stream looks random.

DES was recertified in 1993 by NIST, but its days were limited since it was vulnerable to attack from modern parallel computers. The coding was broken in 1998. Such brute force attacks are feasible if the end results are worth the effort. International banking is an example of this. A single success out of many tries can yield huge sums of money transferred to a wrong account.

Encryption is frequently viewed as too difficult to depend on for secure, tactical communications. If it has not already, this will change in the future primarily because much of the communication—tactical and otherwise—is becoming digital. Signals that start out as digital are easier to encrypt—modern encryption techniques for electronic signaling normally involve converting analog signals to digital before encryption. The advent of widespread dependencies in commerce will cause the issues in encryption to be quickly overcome.

6.10.5 Pretty Good Privacy

In 1991, Phillip Zimmerman developed *pretty good privacy* (PGP). It is an implementation of several of the above concepts for encryption as an alternative to DES. Public and private asymmetric keys are used to set up sessions, where a new session key is exchanged for each new conversation. After establishing the session key, encryption proceeds symmetrically. PGP uses encryption techniques based on the RSA algorithms, although some latest algorithms also support Diffie-Hellman key formats. It also implements digital signatures. The keys can be very long, with current implementations in the 4,000 bit range. There is no reason to limit the key size, however, and if they are needed, longer keys will be implemented.

6.10.6 Fortezza Encryption System

It has long been recognized that certain of the information within the DoD was not classified but sensitive in that it is undesirable for nonprivileged users to have access to it. Logistics information is an example. The amount of ammunition moving into a theater of operations on a daily basis could be used to indicate plans for an impending offensive operation, for example. This type of information is normally treated as unclassified but it would be beneficial to not have it known to the adversary. Commerce and industry have equivalent requirements for protecting information from disclosure to unauthorized sources.

To address this requirement the National Security Agency (NSA) established the Fortezza program. Like most cryptologic programs, there are two fundamental parts—(1) the hardware and associated software to perform the function and (2) the management hierarchy that deals with the cryptologic keys.

The hardware development resulted in a PCMCIA configuration for the implementation, although there is no particular hardware constraint in general that limits this. The PCMCIA architecture has been an industry standard, particularly for small computers, and implementation of the Fortezza system in this configuration would allow for better migration to the maximum possible number of users.

The key management structure has been implemented so that any authorized person or organization can obtain a key for designated authorities. Lists are maintained that are readily available for the holders of these keys.

The Fortezza system implements a two-tiered cryptologic system. Both asymmetric and symmetric approaches are possible within the Fortezza system. In the former, different cryptologic keys are used at either end of the link whereas in the latter the same key is used. The cryptologic keys, which are used to encrypt the message traffic, are first transferred between the two ends of the link using the asymmetric public key encryption scheme described above. Once the cryptologic keys are exchanged, this key is used to encrypt the message traffic symmetrically because of the extensive overhead to perform the calculations for the public key encryption system.

The digital signature standard (DSS) is based on the digital signature algorithm (DSA). It is built into the Fortezza PCMCIA chips.

6.10.7 Escrow Encryption System

A notion for telephone security is the key escrow system proposed in the United States. One implementation of this puts a device unique key, which is part of a chip, called the clipper chip, into telephones after manufacture, but prior to going into operation. When put into operation, the key is split into two parts. Each part of the key is given to two different escrow holders. This approach was initially developed to encrypt voice as might be transmitted over telephones, but it is extendable to data communications as well.

Key escrow systems were proposed to allow the government access to private communications as allowed by wiretapping under the law. Without the ability to perform wiretaps, it is said that illegal activity could be hidden that is now accessible to law enforcement. There is considerable controversy about whether the government should be allowed control over encrypted private communications. In this standard there is a *law enforcement access field* (LEAF), which allows legal wiretaps of such encrypted communication. Key escrow systems proposed in the United States use the escrow encryption standard (EES).

The purpose of EES is to preclude unauthorized parties from listening to telephone conversations. Each clipper chip is assigned a unique identification number and key. This information is retained by the government, with the encryption key split into two halves. Two government agencies each holds half the key.

In reality there are three encryption keys involved with every interaction. The session key is unique to each interaction and is generated and used one time in a symmetric mode. The manufacturers of the phones are free to choose the method of generating the session key—it is not part of the EES. The second key is the aforementioned key unique to each clipper chip. It is used to encrypt the session key. The encrypted session key along with the unique identification number of the phone and a computed checksum are then encrypted with the family key, which is common to all clipper chips. This is then sent to the recipient. The LEAF is sent at the beginning of an EES conversation. If a valid LEAF is not received by the recipient, then conversation is not allowed. The LEAF is 128 bits long. Each chip identification is 32 bits long, and the session key is 80 bits long, while the checksum is 16 bits.

The purpose of the EES is to facilitate legitimate wiretaps, authorized by suitable legal officials (court judges, as wiretaps in the United States are authorized today). Such a wiretap would begin by recording the conversation, including the LEAF. Next the LEAF would be decrypted using the family key, which yields the clipper chip identification, the checksum, and the encrypted session key. Based on the chip identification thus obtained, the two halves of the chip key would be obtained from the holding organizations. Using these two halves the session key would be decrypted leading to the ability to decrypt the conversation.

The National Institute of Science and Technology (NIST), is proposed as one of the two escrow agents. The other is proposed to be the Department of Treasury Automated Systems Division.

6.10.8 Over-the-Air Rekeying

One of the difficult problems with any encryption system is rekeying the encryption devices. Large management structures have been established in the past to perform this function. Rekeying is typically performed by personnel that hand-carry the new keying material to the encryption device and load the new key. This, of course, is time-consuming and expensive. It has been tolerated by the military because of its necessity to ensure the necessary security.

It is possible, however, to rekey remote devices over the air from a central location, but there are problems associated with that as well, one of which is how one insures oneself that the remote encryption device is in the hands of whom it is supposed to be. Therefore protections must be built into the keying system in the event of unusual occurrences.

6.11 Concluding Remarks

Several of the more common techniques for communication systems are presented in this chapter. The older schemes of analog communications are rapidly being replaced with digital communication systems. This is because the latter are more spectrally efficient than analog forms and therefore the precious RF spectrum can be used to support more communications. In situations where fiber can be emplaced to facilitate communications, such as in infrastructure-PSTN systems, that is the media of choice since many gigabits per second can be transported over fiber. In many cases fiber or cable cannot be used. Mobile military forces are an example of this as is the situation where the landscape is very hilly and wire cannot be economically put in place. In those situations RF communications over the air are necessary.

In an information-based society, which most of the developed countries in the world are, the respective military forces become information-based as well. In these situations communication of information is critical. Even in tactical situations, data must be exchanged in increasing volumes that require higher data rates. An example of this is the increased reliance on imagery.

The demand for secure communications, primarily driven by e-commerce over the Internet, is forcing encryption technology to become more reliable and simple to use. Modern encryption technology is becoming widely available worldwide. This may force communication EW systems to change their modes of operation in order to provide information from such systems.

References

[1] Torrieri, D. J., *Principles of Secure Communication Systems*, 2nd Ed., Norwood, MA: Artech House, 1992.

[2] Stuber, G. L., *Principles of Mobile Communication*, Boston: Kluwer Academic Publishers, 1996.

[3] Simon, M. K., S. M. Hinedi, and W. C. Lindsey, *Digital Communication Techniques: Signal Design and Detection*, Upper Saddle River, NJ: Prentice Hall, 1995.

[4] Gagliardi, R. M., *Introduction to Communication Engineering*, 2nd Ed., New York: Wiley, 1988.

[5] Proakis, J. G., *Digital Communications*, 2nd Ed., New York: McGraw-Hill, 1989.

[6] Das, J., S. K. Mullick, and P. K. Chatterlee, *Principles of Digital Communication: Signal Representation, Detection, Estimation, and Information Coding*, New York: Wiley, 1986.

[7] Xiong, F., "Modern Techniques in Satellite Communications," *IEEE Communications Magazine*, August 1994, pp. 84–98.

[8] Xiong, F., "Modern Techniques in Satellite Communications," *IEEE Communications Magazine*, August 1994, p. 89.

[9] Xiong, F., "Modern Techniques in Satellite Communications," *IEEE Communications Magazine*, August 1994, p. 88.

[10] Lathi, B. P., *Communication Systems*, New York: John Wiley & Sons, 1968, p. 228.

[11] Gagliardi, R. M., *Introduction to Communication Engineering*, 2nd Ed., New York: Wiley, 1988, p. 336.

[12] Simon, M. K., et al., *Spread Spectrum Communications*, Volume 1, Rockville, MD: Computer Science Press, 1985, p. 143.

[13] Yost, R. A., and R. H. Pettit, "Susceptibility of DS/FH, Binary DPSK to Partial and Full Band Barrage Jamming," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. AES-17, No. 5, September 1981, pp. 693–700.

[14] Proakis, J. G., *Digital Communications, 3rd Ed.*, New York: McGraw-Hill, 1995, p. 735.

[15] Kohno, R., R. Meidan, and L. B. Milstein, "Spread Spectrum Access Methods for Wireless Communications," *IEEE Communications Magazine*, January 1995, pp. 58–67.

[16] Anderson, T., "HARC: The Newest Wave in Image Compression," *Technology Transfer Business*, Fall 1995, p. 73.

[17] Das, J., S. K. Mullick, and P. K. Chatterlee, *Principles of Digital Communication: Signal Representation, Detection, Estimation, and Information Coding*, New York: Wiley, 1986, p. 505.

[18] Liu, H., H. Ma, M. El Zarki, and S. Gupta, "Error Control Schemes for Networks: An Overview," *Mobile Networks and Applications*, No. 2, 1997, pp. 167-182.

[19] Polydoros, A., and G. Paparisto, "Per-Survivor Processing for Joint Data/Channel Estimation in Multipath Fading and Co-Channel Interference Channels," Communication Sciences Institute, University of Southern California, Electrical Engineering Systems, Los Angeles, CA, 1996.

[20] Divsalar, D., and F. Pollara, "Multiple Turbo Codes," Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, 1995, Figure 1.

[21] Divsalar, D., and F. Pollara, "Turbo Trellis Coded Modulation with Interactive Decoding for Mobile Satellite Communications," Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, 1998.

[22] Poisel, R. A., *Foundations of Communication Electronic Warfare*, Norwood, MA: Artech House, 2008, Ch. 2.

[23] Thompson, B., "Error Codes Widen Design Window," accessed February 2008, http://www.eetimes.com/story/OEG19981231S0003.

[24] Gagliardi, R. M., *Introduction to Communication Engineering*, 2nd Ed., New York: Wiley, 1988, p. 370.

[25] Frater, M. R., and M. Ryan, *Electronic Warfare for the Digitized Battlefield*, Norwood, MA: Artech House, 2001, p. 70.

[26] Fitzgerald, P., "The Quest for Intruder-Proof Computer Systems," *IEEE Spectrum*, July 1989, p. 25.

Chapter 7

Signal Processing

7.1 Introduction

The processing of signals within a communication EW system comes in a variety of forms. Received signals must be processed to extract information from them. Signals to be transmitted from a communication EW system must be processed to be put in the correct format. Thus, understanding the basics of signal processing is important to understand the design and operation of such systems.

Normally signals convey information, thus the interest in signal processing for IW. The light that impinges on our eyes that allows us to see objects are signals. The varying air pressure that carries audio tones detected by our ears are signals. Some characteristic of some medium is changed to facilitate information transfer with signals. The air just mentioned is an example of this. Signal processing is performed to change a signal from one form to another, to impose some characteristic onto a signal, or to facilitate the measurement of some feature of a signal. Features obvious in one domain may be hidden from others, and thus it is necessary to convert the signal to recover such.

When a signal is analog in nature, generally speaking it can be processed in analog form or it can be converted into digital form, in certain cases. An example of an analog signal is the way voice is sent from a telephone—at least prior to the 1990s. Broadcast TV signals in the United States today are analog, although there is a rapid movement to send digital TV signals to the home via cable and direct broadcast satellite.

Some signals originate in digital form, under which circumstances they would have to be converted to analog form in order to process them with analog technologies. Digital signals are normally generated in that form to take advantage of digital signal processing, so conversion to analog is rarely done. Some examples of digital signals being generated that way are random sequences of numbers and digital clocks that display characters rather than sweeping hands.

This chapter is divided into two major parts: the first part discusses some fundamental signal processing principles important to signal processing in EW systems. This includes definition of orthogonal functions, the major transforms usually encountered when processing signals, the theory of sampling bandpass signals, cyclostationary signal processing, and high order statistics. The second part presents some typical applications of signal processing principles including signal detection, classification, and recognition/identification.

7.2 Orthogonal Functions

The set of functions $\Phi = \{\phi_1, \phi_2, \ldots, \phi_N\}$ defined over an interval $x_1 \leq x \leq x_2$, is orthogonal (unitary if the functions are complex) if the following is satisfied

$$\int_{x_1}^{x_2} \phi_i(x)\phi_j^*(x)dx = k_{ij}\delta_{ij} \tag{7.1}$$

where δ_{ij} is the Dirac impulse function defined as

$$\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & \text{otherwise} \end{cases} \tag{7.2}$$

k_{ij} is some constant, and * is complex conjugate. If $k_{ij} = 1$, then the set Φ is said to be orthonormal. Because the results can be easily scaled to or from the unit interval on the real line, it is typically assumed without loss of generality that $x_1 = 0$ and $x_2 = 1$.

$f(x)$ can be approximated by a linear combination of these functions as

$$\hat{f}(x) = \sum_{i=1}^{N} c_i\phi_i(x) \tag{7.3}$$

where

$$c_i = \int_0^1 f(x)\phi_i^* dx \tag{7.4}$$

Φ is thus a set of *basis functions* for $f(x)$. This linear combination representation for $f(x)$ is said to be a *projection* of $f(x)$ onto Φ. Φ is said to be *complete* if any (piecewise) continuous function $f(x)$ can be represented this way and that the *mean square error* (MSE) given by

$$\mathrm{MSE} = \int_0^1 \left| f(x) - \hat{f}(x) \right|^2 dx \qquad (7.5)$$

converges to zero for some N.

When $f(x)$ is defined only at discrete points, then the above integrals are replaced by summations. *Orthonormality* is when

$$\frac{1}{N}\sum_{j=1}^{N} \phi_i(x_j)\phi(x_j) = \delta_{ik} \qquad (7.6)$$

and the approximation function $\hat{f}(x_j)$ is given by

$$\hat{f}(x_j) = \sum_{i=1}^{N} c_i \phi_i(x_j) \qquad (7.7)$$

where

$$c_i = \frac{1}{N}\sum f(x_j)\phi_i^{*}(x_j) \qquad (7.8)$$

Parseval's relationship says that the energy in the x-domain must be equal to the energy in the transformed domain. That is,

$$\sum_{i=1}^{N} |c_i|^2 = \frac{1}{N}\sum_{i=1}^{N} |f(x_i)|^2 \qquad (7.9)$$

7.3 Transforms

A transform is a mathematical manipulation of a signal to convert it from one mathematical domain to another (say from time, t, to frequency, f). There are many ways to change the representation of a signal from one form to another. Such transformations are performed for a variety of reasons. One of these is to ascertain the frequency content of the signal. Another is to compute more efficient ways to transmit the signal from one place to another. Many transforms have been discovered that are useful for processing signals. Only a few of the many transforms will be discussed here.

7.3.1 Trigonometric Transforms

Several transform relationships are based on the trigonometric functions, in particular the cosine and sine functions. Some of these that are important for understanding EW systems are presented in this section.

7.3.1.1 Fourier Transform

One of the primary functions that most RF EW systems perform in signal processing is detecting the presence of signals while simultaneously determining the frequency of these signals. This is frequently accomplished by calculating the Fourier transform of the signals. One of the things for which this transform is useful is determining the energy versus frequency content of a spectrum, and therefore can give an indication of the center frequency as well as a signal's bandwidth.

Perhaps the most prolific transform in signal processing is the Fourier transform, and in particular, the *fast Fourier transform* (FFT). Cooley and Tukey discovered the FFT in 1965 at Bell Laboratories as a fast method of computing the *short-time Fourier transform* (STFT) of a signal on a digital computer. Compressive receivers discussed in Section 9.3.2.3 are essentially Fourier transform calculators, as is the acousto-optical processor based on a Bragg cell. The digital receiver normally computes the Fourier transform to determine the signal spectrum. The FFT is typically calculated with a digital signal processor, usually implemented on one or more semiconductor chips.

Knowing this frequency content is sometimes useful for a variety of reasons. One simple way to visualize this is to consider a signal that has the full bandwidth of the U.S. FM radio band. That is, all of the signals from 88 to 108 MHz are present. Of course, listening to this signal would be nonsensical since no intelligibility could be ascertained. Taking the Fourier transform of such a signal, however, would reveal all of the FM radio channels that are present. Putting an appropriate filter around one of these channels would allow only that signal to get through and therefore it could be clearly heard.

The basis functions for the Fourier transform are $\sin(x)$ and $\cos(x)$. The Fourier transform is a complete transform. Orthogonality dictates that the inner product of the basis functions must equal zero unless the two functions are the same basis function. Let T denote an integral number of periods of the sine and cosine functions. In this case

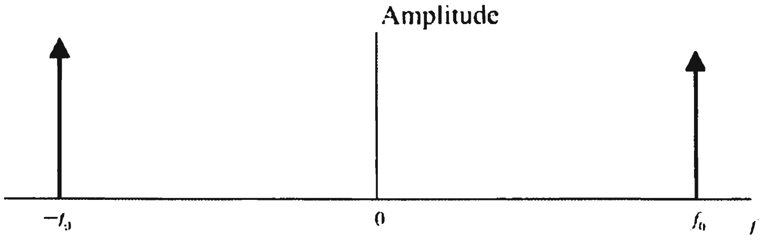


Figure 7.1 Fourier transform of the cosine function.

$$\int_0^T \sin x \cos x dx = \frac{1}{2}\int_0^T \sin 2x dx + \frac{1}{2}\int_0^T \sin(0) dx \tag{7.10}$$

The first integral on the right side equals zero because any sine function integrated over an integral number of periods is zero. The second integral is zero since sin(0) = 0. Thus, the sine and cosine functions are orthogonal.

The Fourier transform of a time signal $s(t)$ is given by

$$S(f) = \int_{-\infty}^{\infty} s(t) e^{-j2\pi ft} dt \tag{7.11}$$

where f is the frequency in hertz, and t is time in seconds. Even though this integral goes from minus infinity to plus infinity, most signals of practical interest have a limited time extent. Some mathematical approximations to real signals do include the unlimited time, however. Limiting the time extent can cause some unexpected results since it is mathematically impossible to simultaneously limit a signal in time and frequency extent. Thus, a signal defined over a time interval (t_1, t_2) has infinite frequency content. Likewise, a signal with frequency extent (f_1,f_2) must exist for all time.

The magnitude of the Fourier transform of an unmodulated cosine wave of frequency f_0 is two impulse functions as shown in Figure 7.1 (phase is ignored for now). The negative frequency might appear odd to some. There is no physical manifestation of negative frequency that is currently known—it is necessary for the mathematics to work out. The Fourier transform is particularly useful for processing smooth, continuous functions. It does not, however, work well with discontinuous functions or nonstationary data. This can be illustrated by trying to approximate a square pulse with sinusoidal functions. The result is known as Gibb's phenomenon, which describes oscillations at the points of discontinuity as shown in Figure 7.2. No matter how many sinusoidal functions are included in the transform, the overshooting oscillations remain. Their amplitude is about 18% of the amplitude of the pulse.

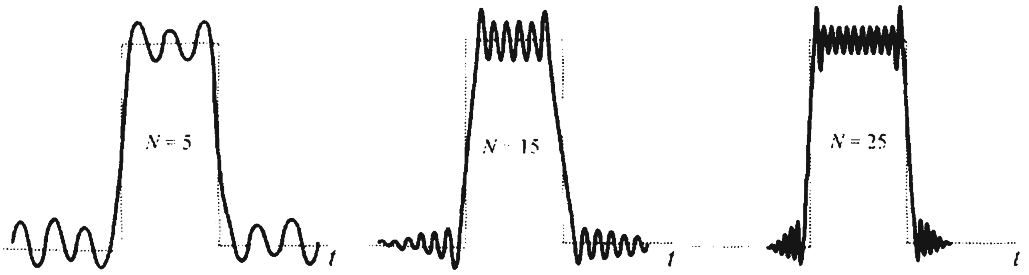


Figure 7.2 Gibb's phenomenon is manifest when trying to approximate a discontinuity with Fourier coefficients.

Since practical implementations of the calculation of the Fourier transform must by necessity cover only a finite amount of time, the STFT was devised. This transform essentially puts a time window around the signal, which has a duration that is shorter than the duration of the whole signal. This calculates the Fourier transform over just this time interval. The effect is to multiply the signal by the unit pulse time window, which has amplitude of 1 and an extent from 0 to T or $-T/2$ to $T/2$ (the same results ensue). Since multiplication in the time domain equates to convolution in the frequency domain, and the spectrum of the unit pulse has a sin x/x shape, the resultant magnitude of the spectrum of the STFT of a sine wave is as shown in Figure 7.3. The spectrum extent (bandwidth) of each impulse is larger than the theoretical single line. The windowing has essentially spread the signal over more frequencies. These frequencies were not part of the original signal, therefore artifacts were introduced into the process.

One of the properties of the STFT is that it is not possible to simultaneously select an arbitrary time window and arbitrary frequency extent. These two parameters are related to each other. The smaller the time resolution, the larger the frequency resolution, and vice versa, the larger the time resolution, the lower the frequency resolution. This is expressed by the uncertainty inequality, which states that in Fourier analysis

$$\Delta f \Delta t \geq \frac{1}{4\pi} \tag{7.12}$$

where Δf and Δt are defined by their second moments divided by their energies of the signal being analyzed.

It is frequently assumed that signals are stationary and ergodic.[1] When a signal is stationary, the Fourier transform statistics do not vary with time since the signal does not. In practice this is rarely true; however, the assumption that the signal is not

[1] A sequence is *stationary* if its statistical properties are constant with time. A sequence is *ergodic* if its statistical properties calculated over time are the same as those calculated across the ensemble of sequences.

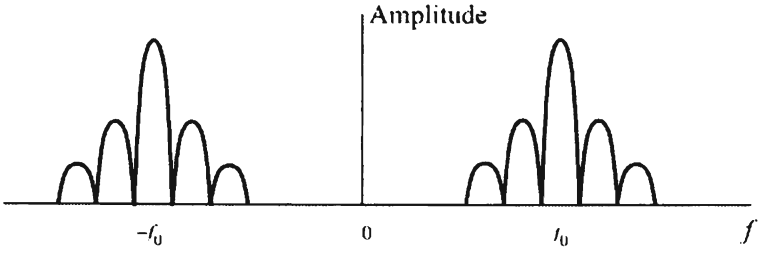


Figure 7.3 Effects of time windowing the cosine function.

changing with time is common. Since this is an STFT, the assumption of stationarity is only necessary over the interval for which data samples of the signal are taken. The ergodic property allows the statistics of a particular realization of a signal from the ensemble to be calculated and used to represent the entire set.

To help deal with nonstationary signals, more complex transformations than the STFT have been devised, such as the Wigner-Ville distribution (the term distribution refers to the distribution of energy with frequency and is irrelevant here). These transforms determine the frequency content versus time. Unfortunately such transforms are nonlinear, and what is called cross-product terms are generated if two or more signals are analyzed simultaneously [1]. The continuous Wigner-Ville transform for signal $s(t)$ is given by

$$W(t,f) = \int_{-\infty}^{\infty} s(t+\tau/2)s^*(t-\tau/2)e^{-j2\pi ft}d\tau \tag{7.13}$$

Thus, the Wigner-Ville distribution is the Fourier transform of the autocorrelation function of the signal. Its discrete form for a time series $x(n)$ is given by

$$W_d(n,f) = 2\sum_{k=-\infty}^{\infty} h_N^2(k)x(n+k)x^*(n-k)e^{-j4\pi fk} \tag{7.14}$$

where $h_N(k)$ is a data window for smoothing the frequency response. One of the advantages of this transform is it is real with no imaginary component as in the STFT.

To help cope with the cross-product problem, the Choi-Williams transform was developed. It is similar to the Wigner-Ville except the cross-product terms are suppressed (but not eliminated). The Choi-Williams transform is given by

$$C(t,f)=\int_{-\infty}^{\infty}e^{-j2\pi ft}\int_{-\infty}^{\infty}\sqrt{\frac{\sigma}{4\pi\tau^2}}e^{-\frac{\sigma(\mu-t)^2}{4\tau^2}}s(\mu+\tau/2)s^*(\mu-\tau/2)d\mu d\tau \qquad (7.\;\;)$$

In this expression σ is a variable that controls the amplitude cross-product terms as well as the frequency resolution. As σ → ∞, this transform becomes the Wigner-Ville distribution.

7.3.1.2 Hartley Transform

The Hartley transform also computes a representation of a signal that displays its energy versus frequency contents. In fact, the Fourier and Hartley transforms are mathematically related to each other. An advantage of the Hartley transform is that the way it can be calculated sometimes is more efficient than the Fourier transform. The Hartley transform works with real numbers, whereas the Fourier transform applies to complex numbers. Another advantage of the Hartley transform is its ease of implementation with real hardware.

The Hartley transform pair is given by

$$H(f)=\int_{-\infty}^{\infty}x(t)\operatorname{cas}2\pi ftdt \tag{7.16}$$

and

$$x(t)=\int_{-\infty}^{\infty}H(f)\operatorname{cas}2\pi ftdf \tag{7.17}$$

where cas(x) = cos(x) + sin(x) (cosine and sine). These are similar to the Fourier transforms but are not the same. For example, $H(f)$ is real where, generally, the Fourier transform is complex [2].

When the input sequence is given by x_k, the discrete Hartley transform is given by the relations

$$H_{N,f}(x)=\frac{1}{N}\sum_{k=0}^{N-1}x_k\operatorname{cas}\left(\frac{2\pi}{N}fk\right) \tag{7.18}$$

and

$$H_{N,k}^{-1}(\nu)=\sum_{f=0}^{N-1}\nu_f \operatorname{cas}\left(\frac{2\pi}{N}fk\right)=x_k \tag{7.19}$$

where, as above, cas(x) = cos(x) + sin(x). The Hartley transform can be calculated with a recurrence relationship which facilitates the fast Hartley transform. This relationship is

$$H_k(\nu)=H_{k-1,\text{odd}}(\nu)\cos\left(\frac{2\pi}{N}\nu\right)+H_{k-1,\text{even}}\sin\left(\frac{2\pi}{N}\nu\right) \tag{7.20}$$

The discrete Hartley transform is quite fast with relatively low computational requirements, especially if the cas(x) terms are precomputed and stored in a table.

7.3.1.3 Discrete Cosine Transform

The DCT is a popular transform in modern image compression algorithms such as JPEG and MPEG. The reason for this is the relatively light computational load involved. The one-dimensional transform relationships are given by

$$X_n=\frac{C(n)}{2}\sum_{k=0}^{N-1}x_k\cos\left[\frac{(2k+1)n\pi}{2N}\right] \tag{7.21}$$

and

$$x_k=\sum_{n=0}^{N-1}\frac{C(n)}{2}X_n\cos\left[\frac{(2k+1)n\pi}{2N}\right] \tag{7.22}$$

where

$$C(n)=\begin{cases}\frac{1}{\sqrt{2}} & n=0\\ 1 & \text{otherwise}\end{cases} \tag{7.23}$$

For transformation and exchange of images, the two-dimensional discrete cosine transform is used. The transform coefficients are determined and those are transmitted versus the raw image samples themselves. This is a more efficient way of transmitting the images. The two-dimensional transform and its inverse are given by

$$X_{k,m} = \frac{2}{N}C(k)C(m)\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} x_{i,j} \cos\left[\frac{(2i+1)k\pi}{2N}\right]\cos\left[\frac{(2j+1)m\pi}{2N}\right] \tag{7.24}$$

and

$$x_{i,j} = \frac{2}{N}\sum_{k=0}^{N-1}\sum_{m=0}^{N-1} C(k)C(m)X_{k,m} \cos\left[\frac{(2i+1)k\pi}{2N}\right]\cos\left[\frac{(2j+1)m\pi}{2N}\right] \tag{7.25}$$

For image processing, these two-dimensional transforms are taken over an $N \times N$ block of pixels denoted by $x_{i,j}$.

There is a fast version of the discrete cosine transform available. Furthermore, this transform is separable in that one-dimensional transforms can be obtained for the rows of the image and then down the columns of the image. Doing so reduces the number of computations involved.

7.3.2 Haar Transform

The Haar transform is useful for processing functions with discontinuities and functions that are nonstationary. It is a complete transform and is one of the fastest available because the only numbers involved are 0, 1, -1, $\sqrt{2}$, and $-\sqrt{2}$. As an example, the 8 × 8 Haar transform matrix is given by

$$\mathbf{H}_8 = \frac{1}{\sqrt{8}}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ \sqrt{2} & \sqrt{2} & -\sqrt{2} & -\sqrt{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sqrt{2} & \sqrt{2} & -\sqrt{2} & -\sqrt{2} \\ 2 & -2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & -2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & -2 \end{bmatrix} \tag{7.26}$$

When the Haar matrix multiplies a signal vector, the signal vector is sampled from low to high frequencies. That is, if

$$\vec{f}(x) = \mathbf{H}_8\vec{g}(x) \tag{7.27}$$

then

$$\begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \\ f_6 \\ f_7 \\ f_8 \end{bmatrix} = \frac{1}{\sqrt{8}} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ \sqrt{2} & \sqrt{2} & -\sqrt{2} & -\sqrt{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sqrt{2} & \sqrt{2} & -\sqrt{2} & -\sqrt{2} \\ 2 & -2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & -2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & -2 \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \\ g_3 \\ g_4 \\ g_5 \\ g_6 \\ g_7 \\ g_8 \end{bmatrix} \tag{7.28}$$

$$= \frac{1}{\sqrt{8}} \begin{bmatrix} g_1 + g_2 + g_3 + g_4 + g_5 + g_6 + g_7 + g_8 \\ g_1 + g_2 + g_3 + g_4 - g_5 - g_6 - g_7 - g_8 \\ \sqrt{2}g_1 + \sqrt{2}g_2 - \sqrt{2}g_3 - \sqrt{2}g_4 \\ \sqrt{2}g_5 + \sqrt{2}g_6 - \sqrt{2}g_7 - \sqrt{2}g_8 \\ 2g_1 - 2g_2 \\ 2g_3 - 2g_4 \\ 2g_5 - 2g_6 \\ 2g_7 - 2g_8 \end{bmatrix} \tag{7.29}$$

Thus, f_1 is a measure of the average, or mean, value; f_2 is a measure of the differences of the means of the first four samples compared to the last four; f_3 is a measure of the differences of the mean values of the first two samples compared with the second two; and so on. The higher the index on f_i, the higher the "frequency" of the measurement of the data. If the input samples, g_i, correspond to a set of pixels from an image, the 8 × 8 Haar transformation produces information about the makeup of the set of pixels. The average value would indicate the brightness of the set, while the last four would facilitate edge detection, for example.

7.3.3 Wavelet Transforms

Wavelet transforms are linear transforms as is the STFT. The transforms deal with the problem of selecting time and frequency windows by using short time windows for high frequencies and long time windows for lower frequencies. The Fourier transform assumes that the signal has been present in its current form forever in the past (stationary), and when it is applied to signals that do not conform to this assumption, then errors occur. Wavelets avoid this situation by not making such an assumption. Therefore, wavelets are useful for nonstationary signal analysis. In particular, they are useful for analysis of signals with discontinuities (or almost discontinuities) in them. A discontinuity is an instantaneous change in a signal that is abrupt and not smooth. Discontinuities do not really exist in real situations, but there are cases where discontinuities almost occur—close enough that the engineering analysis of the signals must treat them as discontinuous. Therefore, the ability to transform such signals is important [3].

Variable-sized time windows are used in wavelet analysis. Short windows are used at high frequencies and long windows at low frequencies. This is like maintaining a constant Q in filters, where the fractional bandwidth is maintained relative to the center frequency. The result of wavelet transforming is time-scale representation versus the time-frequency of Fourier analysis. In the case of Fourier analysis to obtain a realistic representation of the signal in the frequency domain, typically many terms of the transform are required—that is, many of the sine and cosine terms need to be retained. In wavelet analysis typically many fewer terms need to be maintained.

Another advantage of the wavelet transform over the Fourier transform is that the time and frequency resolution of the latter cannot be independently set, whereas in the former they can be. In Fourier analysis, if one wants fine frequency resolution, then time resolution must be sacrificed. If one wants fine time resolution, then frequency resolution must be given up.

Once a window has been selected for the STFT, the time and frequency resolutions remain fixed. For continuous wavelet analysis, it is the ratio $\Delta f / f$ that remains constant, or the frequency resolution is proportional to the frequency. Time resolution becomes arbitrarily good at high frequencies and frequency resolution becomes arbitrarily good at low frequencies.

Shown in Figure 7.4 is the wavelet composition for the sawtooth signal shown at the top [4]. Notice how the basis functions are associated with different portions of the wave. At the discontinuity, the short wavelets are large, while during the longer duration ramps the long wavelets have a larger amplitude. Noise reduction (filtering) is possible using wavelets as shown in Figure 7.5 [4]. This is useful in many areas in communications where signals are corrupted by noise either intentionally or unintentionally. This noise reduction is accomplished by zeroing the basis function coefficients when they are below a certain threshold value. The signal is then reobtained by inverting the transform. Wavelets have found application in characterizing acoustic signals for computer synthesis of music [5], coding for communications over digital channels [6, 7], computer vision [8], and graphics [9].

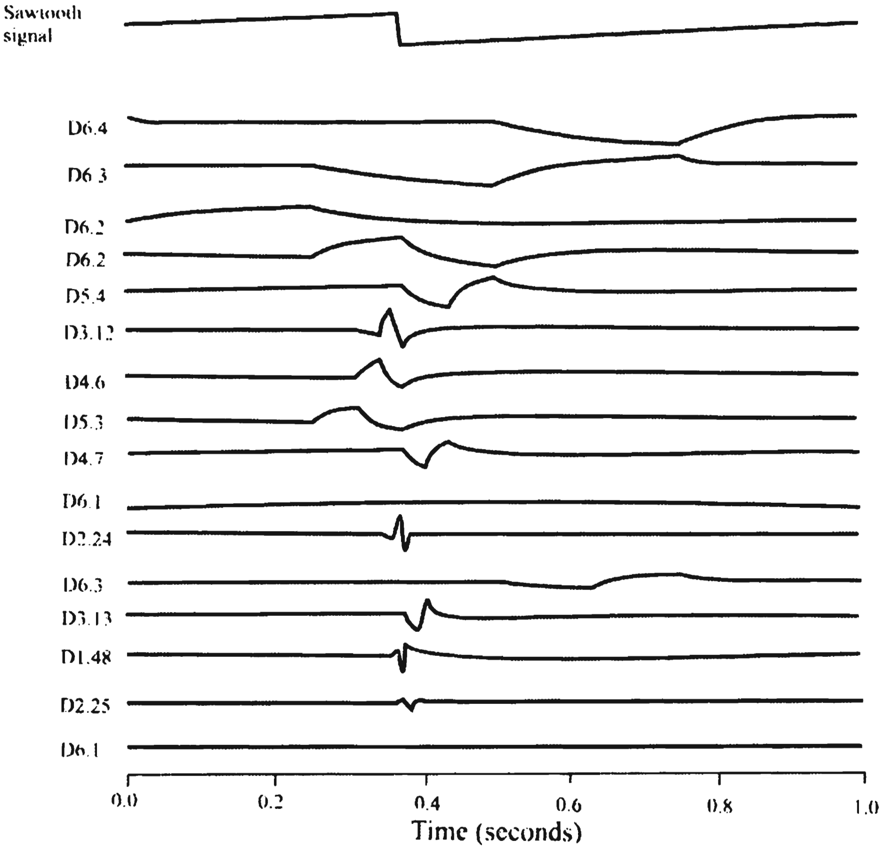


Figure 7.4 Example dilation and scaling of a mother wavelet function. (*Source:* [4]. © IEEE 1996. Reprinted with permission.)

Compression of data is one of the more useful and interesting properties of the wavelet transforms. Most of the energy in image data is located in very few wavelet coefficients. In a typical application, only 3–5% of the wavelet coefficients are necessary to accurately re-create the original data sequence. The other coefficients are set to zero by some mechanism (e.g., thresholds and percentage of coefficients).

If the wavelets that make up the basis are suitably chosen, representation of functions by these coefficients can be very efficient. That is to say, very few of the coefficients are nonzero. Thus, storing the coefficients or transmitting the coefficients rather than the original function can be executed with fewer symbols than the original function.

There are an infinite number of families of wavelets such as the Haar wavelet discussed below. Selection of the right wavelet family to match the problem at hand is one of the challenges of application of the technology. As opposed to the Fourier transform that is defined for an infinite x-scale, wavelets have only compact support. That is to say that they are nonzero for only a segment of the x-axis. The advantage of compact support is the ability of wavelets to represent functions that have

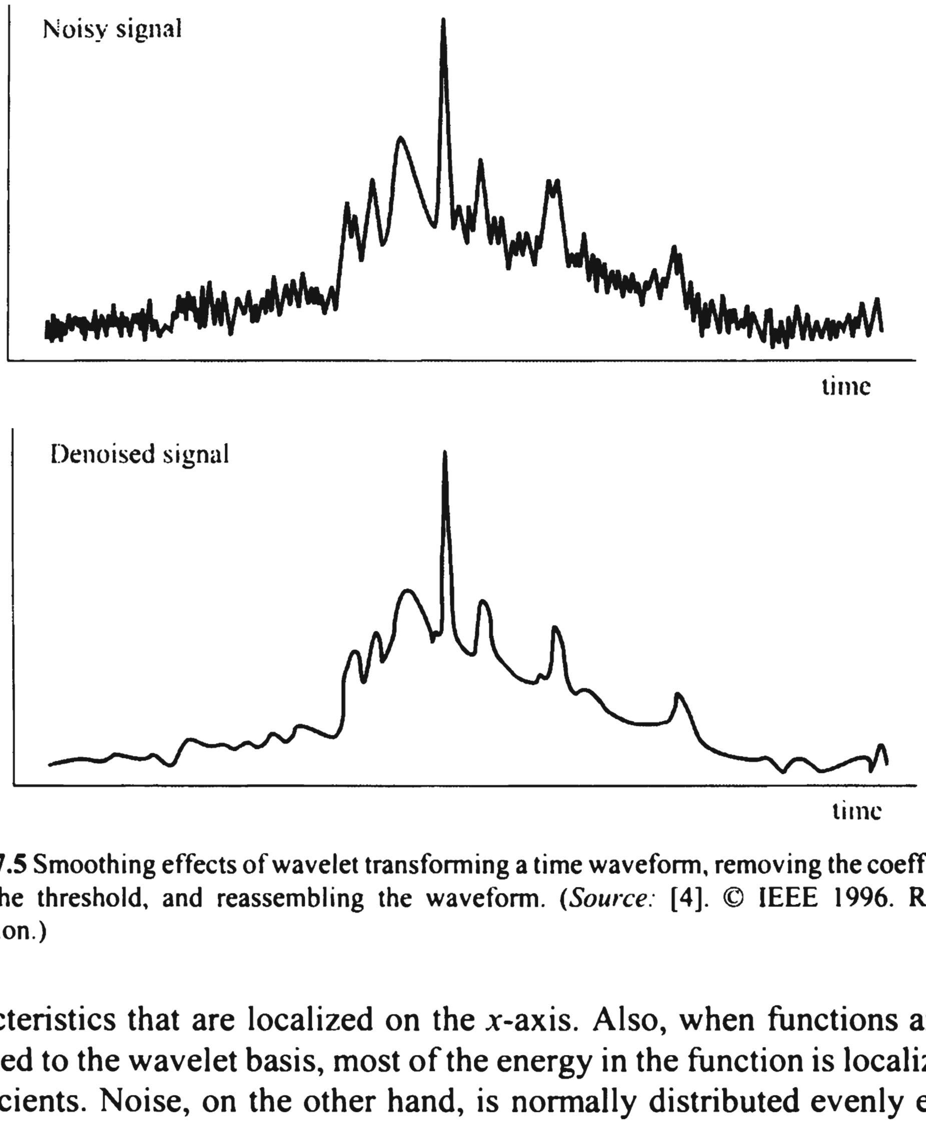


Figure 7.5 Smoothing effects of wavelet transforming a time waveform, removing the coefficients that are below the threshold, and reassembling the waveform. (*Source:* [4]. © IEEE 1996. Reprinted with permission.)

characteristics that are localized on the x-axis. Also, when functions are correctly matched to the wavelet basis, most of the energy in the function is localized to a few coefficients. Noise, on the other hand, is normally distributed evenly everywhere. Thus, to denoise a function, it is only necessary to retain the appropriate few coefficients and zero all the rest. Considerable SNR improvements can ensue from this process.

For the Fourier transform, the basis functions are sine and cosine functions. Wavelets also have basis functions. If $h(t)$ is the prototype wavelet function, also called the mother wavelet, the other wavelet basis functions are given by

$$h_a(t) = \frac{1}{a^{1/2}} h\left(\frac{t}{a}\right) \tag{7.30}$$

where a is a scale factor. Therefore, once the prototype basis function $h(t)$ is specified, the remainder of the basis functions are obtained by expanding and contracting in time and amplitudes as well as by time shifts of the prototype. Thus, the *continuous*

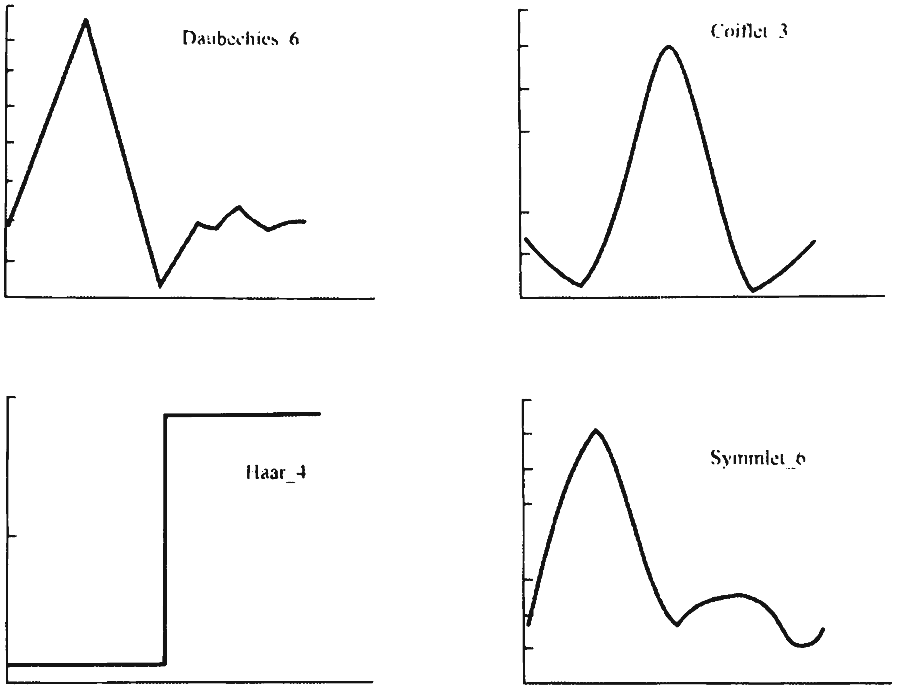


Figure 7.6 Four mother wavelet functions. The other wavelets are scaled and stretched versions of the mother wavelet. (*Source:* [10]. © IEEE 1995. Reprinted with permission.)

wavelet transform (CWT) is given by

$$\mathrm{CWT}_x(\tau,a)=\frac{1}{|a|^{1/2}}\int x(t)h^*\left(\frac{t-\tau}{a}\right)dt \tag{7.31}$$

Shown in Figure 7.6 are four example mother wavelet functions [10]. The _N notations specify the order of the wavelet. Note that just as other basis function sets need not be orthogonal, wavelet families also need not be orthogonal.

The *discrete wavelet transform* (DWT) was discovered by Daubechies [11]. If the mother wavelet is given by $\varphi(x)$, then the DWT of a discrete function $f(i)$ using the wavelet basis functions is given by

$$\mathrm{DWT}(j,k)=2^{-\frac{j}{2}}\sum_{i=0}^{N-1}f(i)\phi(2^{-j}i-k) \tag{7.32}$$

where the basis functions $\varphi(x)$ are recursively given by

$$\varphi(x) = \sum_{i=0}^{M-1} c_i \phi(2x - k) \tag{7.33}$$

where M is the number of nonzero coefficients, also called the order of the wavelet. The inverse transform is given by

$$f(i) = \sum_{j=0}^{N-1} \sum_{k=0}^{M-1} \mathrm{DWT}(j,k) 2^{-\frac{j}{2}} \phi\left(2^{-j} i - k\right) \tag{7.34}$$

These functions form an orthonormal basis. Like the CWT, they are amplitude weighted and shifted versions of the mother wavelet, where the weighting and shifting are powers of two. In this expression j and k are integer coefficients that scale and dilate the mother function. The parameter j determines the width of the wavelet while k controls its location on the x-axis.

A scaling function $w(x)$ is defined based on the mother wavelet as

$$w(x) = \sum_{k=0}^{N-1} (-1)^k c_{1-k} \phi(2x - k) \tag{7.35}$$

The c_k are called wavelet coefficients. These coefficients satisfy constraints such as

$$\sum_{k=0}^{N-1} c_k = 2 \sum_{k=0}^{N-1} c_k c_{k-2l} = 2\delta_l \tag{7.36}$$

where δ is the impulse function and j is the location index. The latter of these two equations is referred to as normalization. The set of coefficients $\{c_0, c_1, \ldots, c_{N-1}\}$ are often thought of as a filter that is applied to the raw data. In doing so, intermediate results ensue that represent smoothed data and detailed data.

Any function $g(x)$ can be written as a series expansion, just as in the Fourier transform above, by

$$g(x) = c_0 + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j - 1} c_{j,k} \phi_{j,k}(x) \tag{7.37}$$

for some appropriate coefficients $c_{j,k}$. The wavelet coefficients for some common wavelet families are given in Table 7.1.

The aforementioned Haar transform is a wavelet transform, although wavelets were not invented when Haar discovered his transform. The Haar transform is based

Table 7.1 Wavelet Coefficients for Some Wavelet Families

Wavelet	c_0	c_1	c_2	c_3	c_4	c_5
Haar	1	1				
Daubechies_4	$\frac{1+\sqrt{3}}{4}$	$\frac{3+\sqrt{3}}{4}$	$\frac{3-\sqrt{3}}{4}$	$\frac{1-\sqrt{3}}{4}$		
Daubechies_6	0.332671	0.806801	0.459877	–0.135011	–0.085441	0.035226

on Haar functions shown in Figure 7.7 where the first eight functions are shown. Haar functions are defined over the interval (0, 1). Clearly these functions satisfy the above requirement that they all are a shifted and/or amplitude varied version of the same function.

7.3.3.1 Haar Wavelet

The function shown in Figure 7.8 is a wavelet function from the Haar family of wavelet functions. The family is comprised of this function and all functions made by shifting this function along the x-axis in integer amounts, stretching/dilating the wavelet out by factors of 2^k, and shifting in multiples of 2^m units, where k and m are integers.

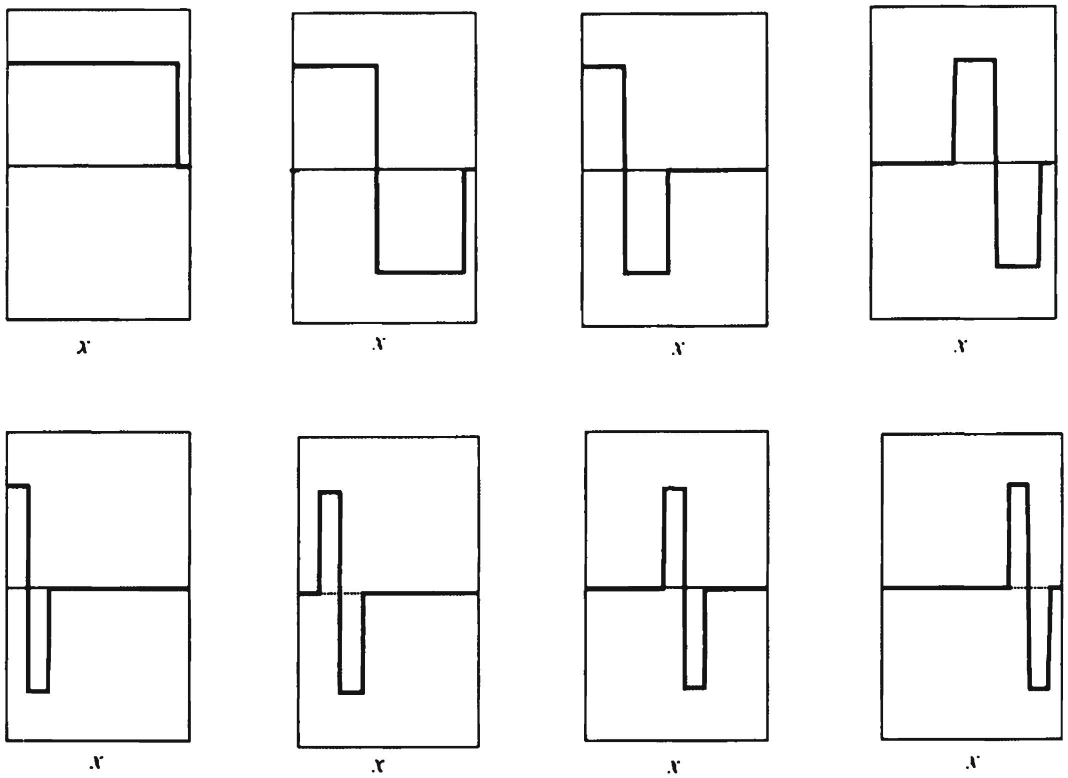


Figure 7.7 Haar functions forming the basis of the Haar wavelet transform.

The dilation shown in Figure 7.8 can be expressed as

$$d(x) = \varphi(2x) + \varphi(2x-1) \tag{7.38}$$

for example. Any dilation can be expressed in this form from the mother wavelet. As is immediately obvious from Figure 7.8, φ(2*x*) has twice the frequency as φ(*x*) and φ(2*x* – 1) is simply an *x*-shifted version (by one place) of φ(2*x*).

The Haar basis functions are defined as follows. The Haar mother wavelet is

$$\varphi(x) = \begin{cases} 1, & 0 \le x < \frac{1}{2} \\ -1, & \frac{1}{2} \le x < 1 \\ 0, & \text{otherwise} \end{cases} \tag{7.39}$$

and

$$\phi_{j,k}(x) = \phi(2^j x - k) \tag{7.40}$$

The functions shown in Figure 7.7 are then

$$\begin{array}{llll} \phi_{0,0} = \phi(x) & \phi_{1,0} = \phi(2x) & \phi_{1,1} = \phi(2x-1) & \phi_{2,0} = \phi(4x) \\ \phi_{2,1} = \phi(4x-1) & \phi_{2,2} = \phi(4x-2) & \phi_{2,3} = \phi(4x-3) & \end{array} \tag{7.41}$$

These functions are orthogonal over (0, 1) because

$$\int_0^1 \phi(x)\phi_{j,k}(x)dx = 0 \tag{7.42}$$

$$\int_0^1 \phi_{j,k}(x)\phi_{n,m}(x)dx = 0 \quad (j,k) \neq (n,m) \tag{7.43}$$

which can be easily verified by examination of Figure 7.7.

The Haar transform is extensively used in image processing, where an image is converted via the Haar transform, and the converted coefficients are transmitted or stored instead of the original image. A fairly good reconstruction of the original image is possible.

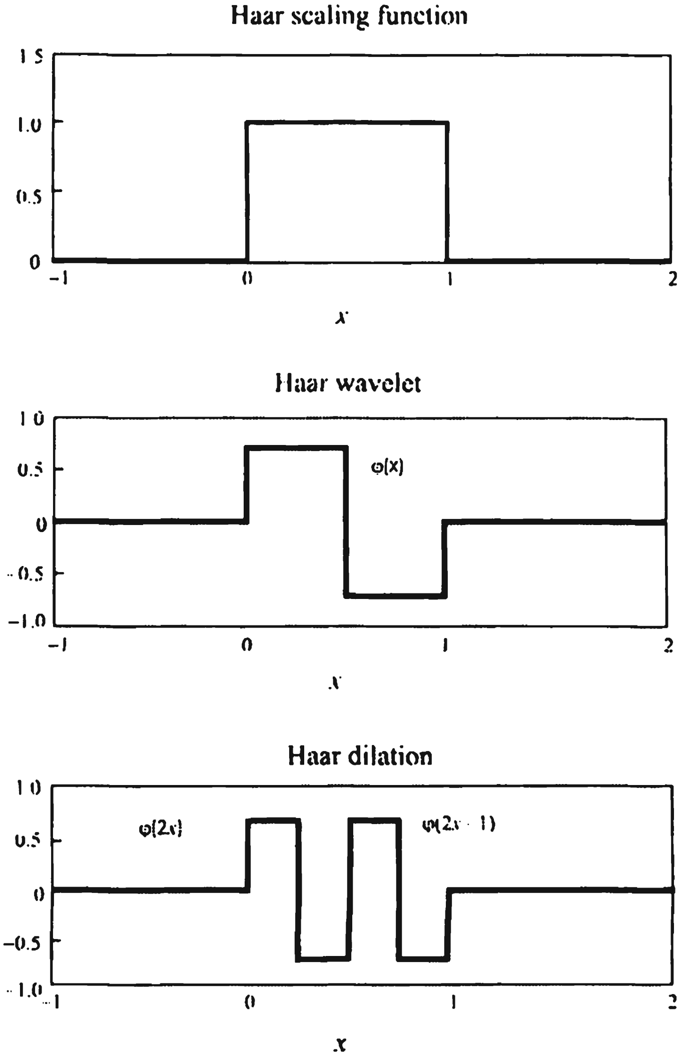


Figure 7.8 The Haar wavelet function.

The Haar transform is particularly applicable to a problem in simultaneous data smoothing and sharpening. Consider the simple case of $N = 4$, and a set of linear equations given by

$$\begin{aligned} y(0) &= \frac{1}{2}[x(0)+x(1)] \quad y(2) = \frac{1}{2}[x(2)+x(3)] \\ y(1) &= \frac{1}{2}[x(0)-x(1)] \quad y(3) = \frac{1}{2}[x(2)-x(3)] \end{aligned} \tag{7.44}$$

In matrix form this is

$$\begin{bmatrix} y(0) \\ y(1) \\ y(2) \\ y(3) \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & -1 \end{bmatrix}\begin{bmatrix} x(0) \\ x(1) \\ x(2) \\ x(3) \end{bmatrix} \tag{7.45}$$

The sum equations perform averaging of two sequential time samples and the difference equations are of the form of a discrete differentiation. The net effect of this transformation is to smooth the data with the sum equations and to sharpen the data differences with the differentiation. The differentiation will tend to enhance sharp transitions in the input data. The sum expressions implement a low-pass filter, which suppresses high-frequency information in the input data, while the difference expressions form a high-pass filter. This filter suppresses low-frequency information while enhancing high-frequency information.

For larger data sets the recursion equations for this problem are given by

$$\begin{aligned} y(2n) &= \frac{1}{2}[x(2n) + x(2n+1)] \quad \text{low-pass filter} \\ y(2n) &= \frac{1}{2}[x(2n) - x(2n+1)] \quad \text{high-pass filter} \end{aligned} \tag{7.46}$$

7.3.4 Fast Transforms

The transforms discussed above, indeed most transforms common in communication EW signal processing, require on the order of N^2 [denoted as $O(N^2)$] operations to complete where N is the transform size. For large values of N, such computations can become unmanageable, especially when the signal processing must be done in real or near-real time. For most of the unitary transforms when N is a power of 2, however, fast versions exist. They are based on factoring the transform matrix into subproblems, and the results of these subproblem computations can be used in subsequent processing without the need to redo them. The fast Fourier transform, for example, can be computed in $O(N\log N)$ computations. A comparison of these quantities is informative. One such comparison for several values of $N = 2^M$ is presented in Table 7.2. Clearly, even for relatively small values of M, the number of operations in the long (nonfast) transforms quickly becomes prohibitive.

The fast transforms are often represented in the form of flow diagrams. As an example, consider the 8 × 8 Haar transform above. Equation (7.26) gives the transform matrix. The flow matrix for this transform is given in Figure 7.9 [12]. The

Table 7.2 Comparison of the Approximate Number of Computations to Compute a Transform

M	$N = 2^M$	N^2	$N \log N$
1	2	4	0.60
2	4	16	2
3	8	64	7
4	16	256	19
5	32	1,024	48
6	64	4,096	116
7	128	16,384	270
8	256	65,536	617
9	512	262,144	1,387
10	1,024	1,048,576	3,083

$1/\sqrt{8}$ is ignored here. This fast transform requires $2(N-1)$ additions and subtractions, the fastest unitary transform available.

7.4 Signal Sampling

Much of the modern signal processing in EW systems is performed in the digital domain. This is due to several reasons; among them are:

- Ease of implementing certain functions;
- Performance stability that does not change with time and temperature;
- Availability of high performance/speed general purpose processors.

The signals to be processed must first be converted into digital form if they do not originate as such. For example, even though the information being conveyed by digital communication signals is inherently digital, the modulated signal is not transmitted digitally. It is modulated onto a suitable carrier signal, frequently modeled as a sinusoidal signal, prior to transmission. This signal is an analog signal.

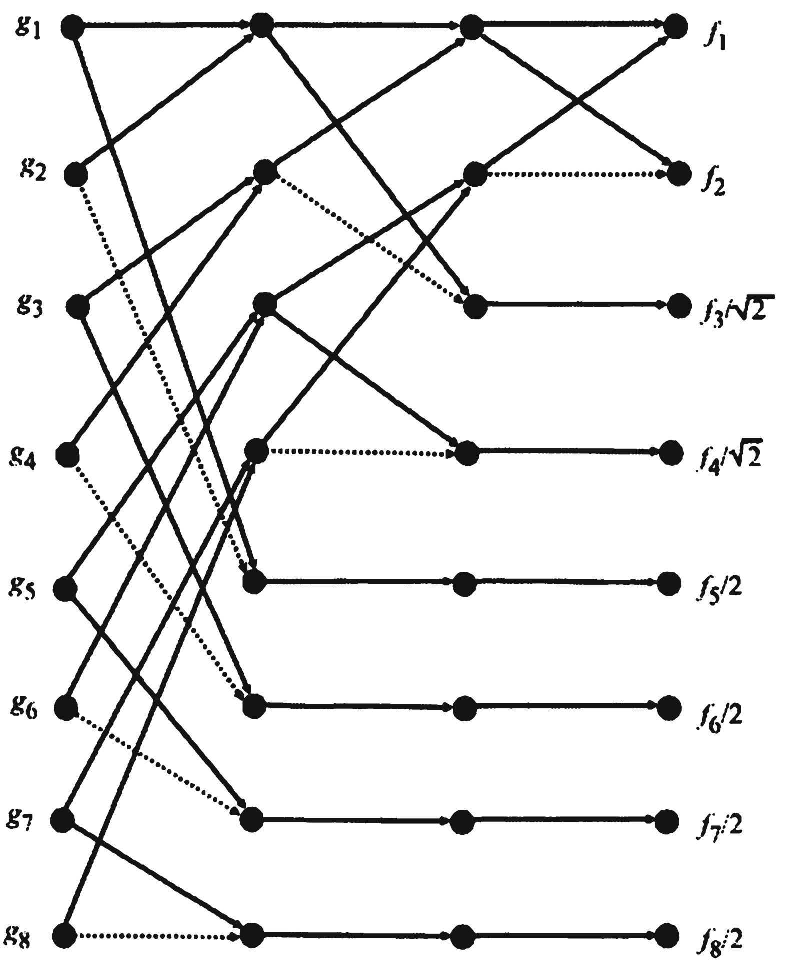


Figure 7.9 Flow diagram for the fast Haar transform. Solid lines are additions while the dashed lines are subtractions. (*Source:* [12]. © R. Bock 1999. Reprinted with permission.)

Conversion of an analog signal into an equivalent[2] digital form requires sampling the analog signal, usually with an ADC. The rate of this sampling has certain requirements imposed on it which are largely determined by the bandwidth, B, of the analog signal. This inherently implies that we are dealing with band-limited signals. The bandwidth is defined as the region between f_l and f_U shown in Figure 7.10. Note that this definition, for reasons that are explained below, does not include f_L or f_U; therefore $B < f_U - f_L$.

7.4.1 Band-Limited Sampling

A signal is *band-limited* if the amplitude of its frequency spectrum is nonzero for only a band of frequencies. Consider signal $g(t)$ whose spectrum is shown in Figure 7.11

[2] Equivalent here means that the original analog signal can be uniquely and perfectly recovered from its equivalent sampled form.

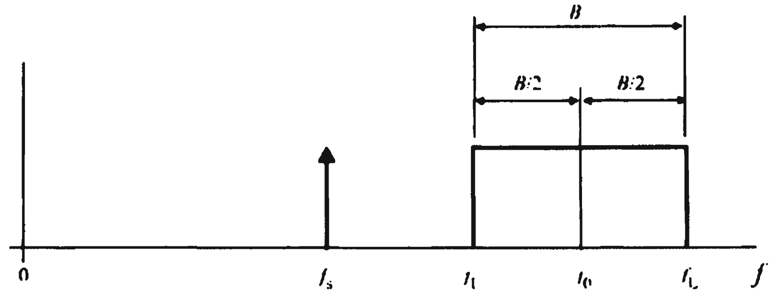


Figure 7.10 Signal bandwidth definitions.

which is zero for frequencies above f_U. For this *baseband* or *low-pass* signal, f_U is also the bandwidth (B). (The bandwidth of a base-band signal is defined only for positive frequencies because the physical significance of negative frequencies has yet to be discovered.) We can represent sampling $g(t)$ in mathematical form by multiplying $g(t)$ by an infinitely long train of impulse functions. Although not necessary, uniform sampling is assumed here for ease of explanation. Uniform sampling is characterized with a single sampling time interval denoted by T. These impulse functions are expressed mathematically by

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

Note that the sampling frequency is given by

$$f_{\text{sampling}} \triangleq f_s = \frac{1}{T} \tag{7.47}$$

The resulting function after multiplying $g(t)$ by an impulse function has the value of $g(t)$ at the instant of sampling, and is zero elsewhere. That is

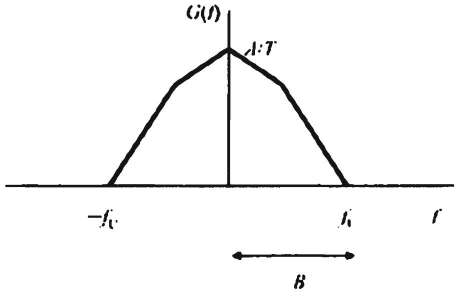


Figure 7.11 Frequency spectrum of the signal, $G(f)$.

$$g(t)\delta(t-nT)=\begin{cases}g(nT), & t=nT\\ 0, & \text{otherwise}\end{cases} \tag{7.48}$$

Based upon (7.48), the sampled signal, denoted $\hat{g}(t)$, is expressed as

$$\hat{g}(t)=g(t)\sum_{n=-\infty}^{\infty}\delta(t-nT) \tag{7.49}$$

The Fourier transform of the sampled signal is given by

$$\hat{G}(f)=\mathcal{F}\{\hat{g}(t)\}=\int_{-\infty}^{\infty}\hat{g}(t)e^{-j2\pi ft}dt \tag{7.50}$$

Recall that multiplication in the time domain corresponds to convolution in the frequency domain. Thus $\hat{G}(f)$ can be expressed as

$$\hat{G}(f)=G(f)*\mathcal{F}\left\{\sum_{n=-\infty}^{\infty}\delta(t-nT)\right\} \tag{7.51}$$

Since we know the spectrum of the original signal $G(f)$ given in Figure 7.11, we need to find the Fourier transform of the train of impulses to evaluate (7.51) and that is what we will do next.

The train of impulses is a periodic function and can, therefore, be represented by a Fourier series as

$$\sum_{n=-\infty}^{\infty}\delta(t-nT)=\sum_{n=-\infty}^{\infty}a_n e^{j2\pi\frac{n}{T}t} \tag{7.52}$$

where the Fourier coefficients are given by

$$a_n=\frac{1}{T}\int_{-T/2}^{T/2}\sum_{n=-\infty}^{\infty}\delta(t-nT)e^{-j2\pi\frac{n}{T}t}dt \tag{7.53}$$

The summation in (7.53) consists of an infinite number of impulse functions. However, the limits of the integral which includes those impulse functions are (*−T*/2, *T*/2). The only nonzero element of the series of impulse functions in this interval is the one at $n = 0$. Therefore the integration produces the same result if the limits on the integral are changed to $(-\infty, \infty)$ and the summation is changed to include only the nonzero component at $n = 0$. The impulse function at $n = 0$ is just $\delta(t - 0t) = \delta(t)$. Thus, we can rewrite (7.53) as

$$a_n = \frac{1}{T}\int_{-\infty}^{\infty} \delta(t) e^{-j2\pi\frac{n}{T}t} dt = \frac{1}{T}\mathcal{F}\{\delta(t)\}\Big|_{f=n/T} = \frac{1}{T} \tag{7.54}$$

since[3] $\mathcal{F}\{\delta(t)\}\big|_{f=n/T} = 1$, and then (7.52) becomes

$$\sum_{n=-\infty}^{\infty} \delta(t - nT) = \frac{1}{T}\sum_{n=-\infty}^{\infty} e^{j2\pi\frac{n}{T}t} \tag{7.55}$$

A signal in the time domain can be obtained from its inverse Fourier transform as

$$f(t) = \mathcal{F}^{-1}\{F(f)\} \triangleq \int_{-\infty}^{\infty} F(f) e^{j2\pi ft} df \tag{7.56}$$

Using the sampling property of the impulse function we get

$$\mathcal{F}^{-1}\{\delta(f - f_0)\} = \int_{-\infty}^{\infty} \delta(f - f_0) e^{j2\pi ft} df = e^{j2\pi f_0 t} \tag{7.57}$$

and therefore

$$e^{j2\pi\frac{n}{T}t} = \int_{-\infty}^{\infty} \delta\left(f - \frac{n}{T}\right) e^{j2\pi ft} df \tag{7.58}$$

$$\mathcal{F}\left\{\sum_{n=-\infty}^{\infty} \delta(t - nT)\right\} = \frac{1}{T}\sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{T}\right) = \frac{1}{T}\sum_{n=-\infty}^{\infty} \delta(f - nf_s) \tag{7.59}$$

[3] This result also follows from the sampling property of the impulse function. That property says

$$\int_{-\infty}^{\infty} \delta(t - 0) e^{-j2\pi\frac{n}{T}t} dt = e^{-j2\pi\frac{n}{T}t}\Bigg|_{t=0} = 1$$

A good overview of the impulse function along with several of its properties is given by Brigham [13].

Based on (7.59) the Fourier transform of the sampled function can be written as

$$\hat{G}(f) = G(f) * \frac{1}{T}\sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{T}\right) \tag{7.60}$$

The convolution of two functions $A(f)$ and $B(f)$ is defined as

$$A(f) * B(f) \triangleq \int_{-\infty}^{\infty} A(z)B(f-z)dz = \int_{-\infty}^{\infty} A(f-z)B(z)dz \tag{7.61}$$

and so

$$\hat{G}(f) = \int_{-\infty}^{\infty} G(f)\frac{1}{T}\sum_{n=-\infty}^{\infty} \delta\left(f - z - \frac{n}{T}\right)dz$$

but an integral is simply the limiting form of a summation so the order of the integration and summation can be exchanged yielding

$$\hat{G}(f) = \frac{1}{T}\sum_{n=-\infty}^{\infty} \int_{-\infty}^{\infty} G(f)\delta\left(f - z - \frac{n}{T}\right)dz$$

Again, using the sampling property of the impulse function:

$$\begin{aligned}\hat{G}(f) &= \frac{1}{T}\sum_{n=-\infty}^{\infty} G\left(f - \frac{n}{T}\right) \\ &= \frac{1}{T}\sum_{n=-\infty}^{\infty} G(f - nf_s)\end{aligned} \tag{7.62}$$

Equation (7.62) is commonly called the *sampling theorem*. It shows that sampling in the time domain at uniform intervals of T seconds replicates the spectrum of our unsampled signal every $1/T$ Hz. This is illustrated in Figure 7.12, which shows the spectrum of $\hat{G}(f)$. The peak amplitude of each nonzero segment of the spectrum has been scaled by $1/T$ to A/T as shown.

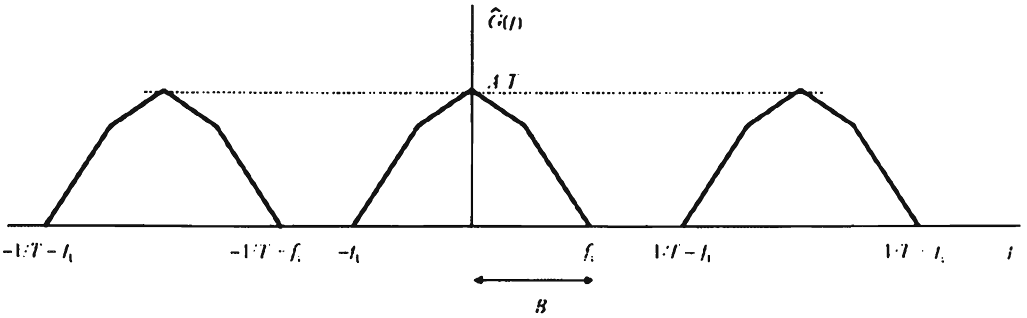


Figure 7.12 Frequency spectrum of the sampled signal consists of $G(f)$ repeated at multiples of the sampling frequency.

7.4.2 Aliasing

As illustrated in Figure 7.12, the individual spectrum copies do not overlap. This is a requirement if the original signal is to be recovered after manipulation in the frequency domain. *Aliasing* is the appellation applied when two spectra replicas overlap. The spectra combine and new frequency components are generated. Artificial signals are "aliased" into the original spectra. This is shown in Figure 7.13, where a high-frequency component f_H has been aliased into the baseband spectrum. From Figure 7.12, to avoid aliasing, we must ensure that $1/T > 2f_U$, or $1/T > 2B$. This result can be expressed in terms of the sampling frequency as

$$f_s \geq 2B \tag{7.63}$$

This result is known as the *Nyquist criterion*. The signal can be recovered from its sampled version by using a low-pass filter to isolate the original spectrum and by attenuating everything else. Thus, extracting the signal with a low-pass filter of cutoff frequency f_U does not eliminate the aliased high frequency.

Consider a special class of band-limited signals known as bandpass signals. A bandpass signal is characterized by $f_L > 0$. For example, the bandpass signal shown

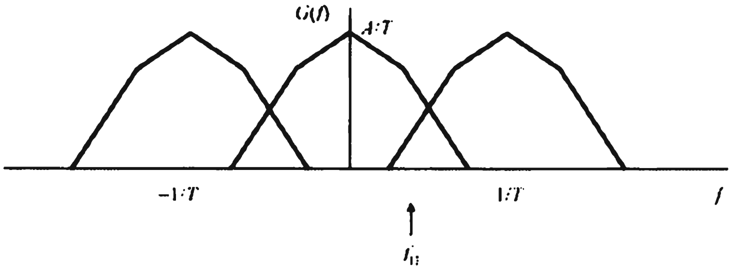


Figure 7.13 Aliasing. The sampling frequency is not high enough to prevent overlap of the repeated spectra of $G(f)$.

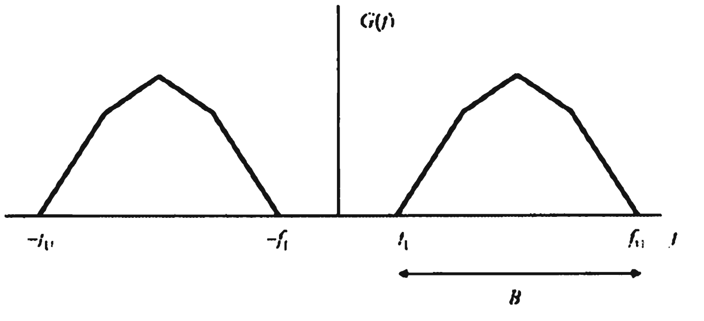


Figure 7.14 Frequency spectrum of bandpass signal $G(f)$.

in Figure 7.14 has signal energy between the frequencies f_L and f_U, and its bandwidth is defined as $B = f_U - f_L$.

As determined above, by sampling this signal at uniform sampling rate T we replicate its spectrum at intervals of $1/T$. Because that spectrum includes a substantial zero-amplitude band between zero and f_L, the actual signal bandwidth is smaller than f_U. We can, therefore, use a sampling frequency lower than that required for a signal whose spectrum occupies all frequencies from $f_L = 0$ to f_U. For example, assume $B = f_U/2$. To satisfy the Nyquist criterion, our minimum sampling frequency is f_U, producing the sampled-signal spectrum of Figure 7.15.

We see that this sampling produces no aliasing, so we could extract the original signal from the samples with a perfect bandpass filter. It is important to note in this example the difference between a baseband signal and a bandpass signal. For baseband signals, the bandwidth, and hence the sampling frequency, depend solely on the highest frequency present. For bandpass signals, bandwidth is usually smaller than the highest frequency.

7.4.3 Recovering Sampled Signal

These characteristics determine the method for recovering the sampled signal: Consider a baseband and a bandpass signal, each with the same value of maximum frequency. The bandpass signal permits a lower sampling frequency only if the method of recovery includes a bandpass filter that isolates the original signal spectrum (the white rectangles shown in Figure 7.16). A low-pass filter (used for baseband recovery) cannot recover the original bandpass signal because it includes the shaded

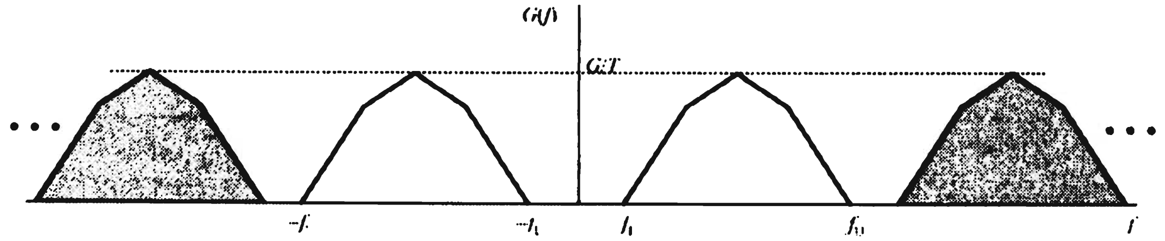


Figure 7.15 Spectrum of sampled bandpass signal again consists of multiple $G(f)$.

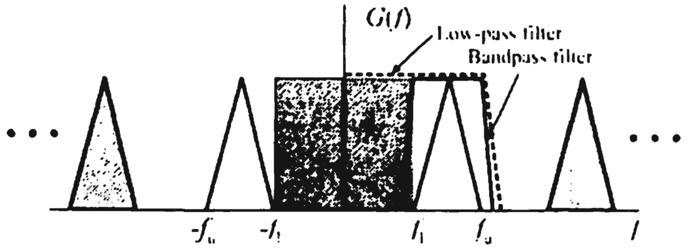


Figure 7.16 Recovery of bandpass signals.

areas shown in Figure 7.16. Thus, if we use a low-pass filter to recover the bandpass signal in Figure 7.16, we must sample at $2f_\mathrm{U}$ to avoid aliasing.

Thus, band-limited signals can be sampled and fully recovered only when observing the Nyquist criterion. For bandpass signals, the Nyquist criterion will ensure no aliasing only when the recovery of the signal is done with a bandpass filter. Otherwise, a higher sampling frequency will be required.

One last consideration is our assumption of band-limited signals. Mathematically, a signal can never be truly band-limited. We cannot simultaneously have a band-limited signal that only exists for a finite period of time. That is, if a signal is finite in time, its spectrum extends to infinite frequency, and if its bandwidth is finite, its duration is infinite in time. Nevertheless, these conditions can be approximated arbitrarily close. The analysis presented here is effective in most practical situations of interest.

The sampling theorem for uniform samples can be generalized to state that the minimum sampling frequency is given by

$$f_{\mathrm{s,min}} = 2\frac{f_\mathrm{U}}{n} \tag{7.64}$$

where

$$n = \left\lfloor \frac{f_\mathrm{U}}{B} \right\rfloor \tag{7.65}$$

is the largest integer in f_U/B. Figure 7.17 illustrates the minimum rate for bandpass filters as represented by (7.64). The theoretical minimum rate $f_\mathrm{s} = 2B$ only applies for integer band positioning.[4] The lines in Figure 7.17 correspond to the minimum

[4] The *band positioning* of bandpass signals is the fractional number of bandwidths from the origin where f_l lies. *Integer band positioning* is when $f_\mathrm{l} = cB$ where c is a non-negative integer. Low-pass corresponds to $c = 0$.

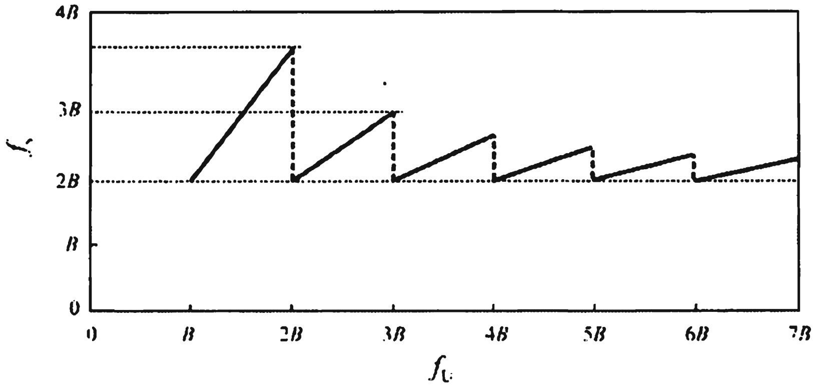


Figure 7.17 Minimum sampling frequency for a bandwidth *B*.(*Source*: [14]. © IEEE 1991. Reprinted with permission.)

uniform rate required for when the band positioning is not an integer. The vertical, dotted lines are discontinuities in the function and are not permitted sampling rates.

Again it should be noted that $2B$ is a valid sampling rate only if there are no signal components at f_L or f_U. If there were energy at those frequencies then they would result in aliasing.

Palpably it makes good engineering sense to include guard bands between the minimum sampling rate and the edges of the pass-band. Otherwise imperfections in implementation and drifts that occur over time and temperature could result in aliasing.

7.5 Cyclostationary Signal Processing

Most analysis of stochastic (random) signals assumes that those signals have stationary properties—that is, the properties do not change with time. This assumption frequently does not apply. In fact, a generalization of the stationarity property for most signals of practical interest is periodic variation of the statistical properties. Such signals are called cyclostationary because their statistical properties are cyclic—they are periodic with time. Computation of the cyclic properties can yield results that can be used for a variety of purposes. One of these is to classify signals by modulation type. Shown in Figure 7.18 [15] are the cyclostationary power spectral densities of a QPSK signal and an MSK signal. Note that while these characteristics for $\alpha = 0$ are essentially the same, for $\alpha \neq 0$ they are dramatically different. For QPSK there are distinctive features at $\alpha = nf_0$ for integer n, whereas for MSK the features are only present for even values of n, and at $\alpha = \pm nf_c \pm nf_0$ for odd values of n. Here f_c is the carrier and f_0 is the baud rate of the digital signal. These differences can be used to discern the type of signal being analyzed. The power spectral density when $\alpha = 0$ corresponds to the normal noncyclic psd that is usually analyzed. This spectrum is

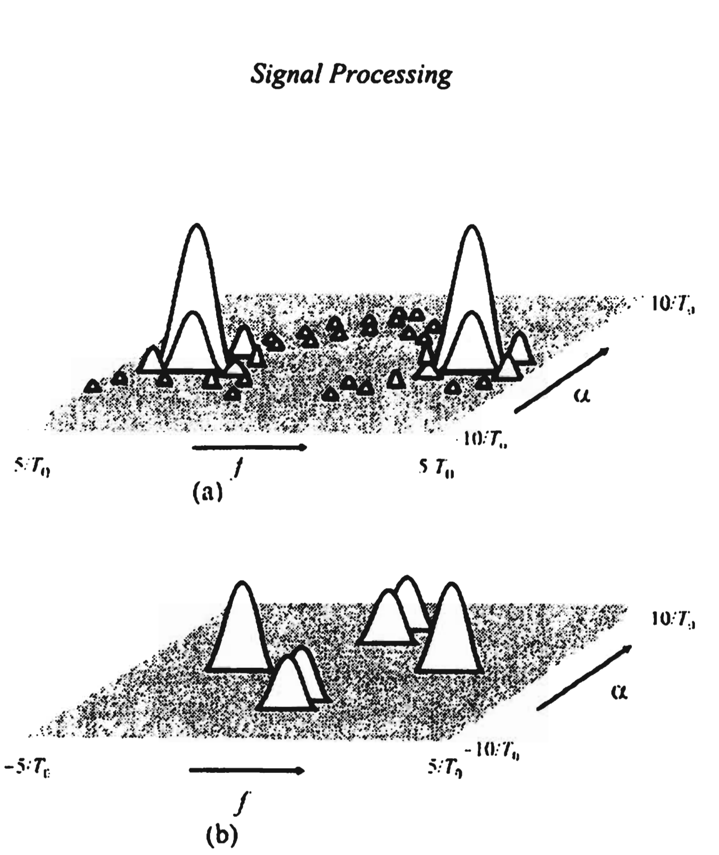


Figure 7.18 Examples of the cyclostationary spectrum of two digital signals: (a) QPSK and (b) MSK. The dramatic difference in form facilitates modulation recognition. The MSK signal in (b) is an SQPSK signal with a raised cosine carrier envelope. (*Source:* [15]. © IEEE 1988. Reprinted with permission.)

unique for this type of modulation and if automated means are available to evaluate this spectra, then the modulation can be uniquely identified.

Some of the uses that have been identified include signal detection and classification, parameter estimation, TDOA estimation, spatial filtering, direction finding, frequency shift filtering for signal extraction, and frequency shift prediction.

In a tactical military setting, identification of the modulation type is extremely useful for a variety of reasons. First, it provides an indication of the possible military unit type with which the radio is associated and this can provide information of the echelon and other relevant information known as the EOB. Second, in those cases where demodulation assets are assigned automatically, determination of the modulation type is first required in order to properly assign the resources. This latter property can be useful in a nonmilitary setting when using general-purpose modems, for example, and it is necessary to know the modulation type in order to select the demodulator.

Another advantage is that Gaussian noise does not exhibit cyclostationary characteristics. Therefore, the cyclostationary spectrum is noise-free and the intercept range of ES systems is thus extended, which equates to increased sensitivity. This noise-free characteristic can be seen in Figure 7.18.

7.6 Higher-Order Statistics

Calculations of the statistical properties of random processes of which most of us are aware are called *first-order statistics*. This notion can be generalized to higher orders by generalizing the equations used to calculate the statistics. If $\mathbf{x} = \{x_i\}$, $i = 1, \ldots, N$, represents random sequences or samples of a signal, then the moments of that sequence are given by

$$m_n = \mathcal{E}\{x^n\} \tag{7.66}$$

where $\mathcal{E}\{\ \}$ represents the familiar expected value operator

$$\mathcal{E}\left\{x^n\right\} = \int_{-\infty}^{\infty} x^n p(x) dx \tag{7.67}$$

and where $p(x)$[5] is the pdf of x [16].

If the signal is ergodic and stationary, then these moments can be estimated with

$$m_n = \frac{1}{N}\sum_{i=0}^{N-1} x_i^n \tag{7.68}$$

If the sequence has zero mean, then the cumulants of order n of that series are given by

$$c_n(t_1, t_2, \ldots, t_n) = \mathcal{E}\{x(t)x(t+t_1)x(t+t_2)\cdots x(t+t_n)\} \tag{7.69}$$

In particular, if $n = 2$, the second-order cumulant of $x(t)$ is

$$c_2(t_1) = \mathcal{E}\{x(t)x(t+t_1)\} \tag{7.70}$$

That is, the second-order cumulant is the autocorrelation function of x. The Fourier transform of $c_2(t_1)$ is the familiar psd function:

$$\text{psd}(f_1) = \int_{-\infty}^{\infty} c_2(t_1) e^{-j2\pi f_1 t} dt_1 \tag{7.71}$$

[5] x^n is shorthand notation for $\left\{x_1^n, x_2^n, \ldots, x_N^n\right\}$.

In general, the Fourier transforms of the higher-order spectra are called polyspectra for obvious reasons.

The third-order cumulant of x, given by

$$c_3(t_1,t_2) = \mathcal{E}\{x(t)x(t+t_1)x(t+t_2)\} \tag{7.72}$$

and its Fourier transform is known as the bispectrum, denoted as B_x

$$B_x(f_1,f_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} c_3(t_1,t_2)e^{-j2\pi(f_1t_1+f_2t_2)}dt_1dt_2 \tag{7.73}$$

Normally a more efficient way to compute B_x is

$$B_x(f_1,f_2) = \mathcal{E}\{X(f_1)X(f_2)X^*(f_1+f_2)\} \tag{7.74}$$

where $X(f)$ is the Fourier transform of $x(t)$.

The first four cumulants can conveniently be calculated using

$$c_1 = m_1 \tag{7.75}$$

$$c_2 = m_2 - m_1^2 \tag{7.76}$$

$$c_3 = m_3 - 3m_1m_2 + 2m_1^3 \tag{7.77}$$

$$c_4 = m_4 - 3m_2^2 - 4m_1m_3 + 12m_2m_1^2 - 6m_1^4 \tag{7.78}$$

These are called *higher order spectra* (HOS).

There are several potentially useful properties of cumulants. If $x(t)$ is a random process with a symmetric pdf, which includes the Gaussian random variable case, then all cumulants higher than the second are equal to zero. In addition, the cumulants of the sum of two random variables are equal to the sum of the cumulants of the individual variables.

If Gaussian noise is added to a non-Gaussian communication signal, then the cumulants of this sum will be those of the signal, since the cumulants of the noise component are all zero. Almost Gaussian noise is generated in the front-end electronics of receiving systems. It is almost Gaussian because, in fact, it is band-limited, whereas true Gaussian noise is not. It is this noise source, in addition to noise from external sources, such as the electrical fields set up by close-by generators, that limit the range of ES systems. Thus, calculating the cumulants of such a time series can extend the dynamic range of EW systems.

Since the higher-order statistics of Gaussian noise are zero, if they are computed for signal detection purposes, there should be no noise present and the detrimental effects of noise are therefore mitigated. In practice, of course, noise is not truly Gaussian and there are usually some residual effects present in the HOS.

Thus, the higher-order statistics associated with a random sequence x can be used for several purposes. They can be used to minimize or eliminate some forms of additive noise and can be used for signal detection in otherwise noisy environments. Some types of signals have features that show up only in higher-order statistical calculations that can be used for signal classification.

7.7 Applications

Some applications of the above generic signal processing techniques in communication EW systems are presented in this section. These include signal detection, signal classification, entity recognition and identification, language identification, and emitter identification. These are but examples of the myriad of signal processing tasks in EW systems.

7.7.1 Signal Detection

In many communication EW systems a receiver is used to search the RF spectrum to seek channels where energy is present. This is the signal detection problem, and it is one of the first functions such systems must perform. Signal detection is frequently combined with signal classification discussed subsequently [17, 18].

Searching the RF spectrum can be further divided into general search and directed search. In general search one or more scan ranges are usually specified in the form of f_{start} to f_{stop}. A frequency step is also sometimes specified, which is usually one channel wide (e.g., in the military VHF frequency range, the channels are 25 kHz wide). Searching then starts at f_{start} and stops at f_{stop}. Frequently this searching is continuous and repetitive. General search is usually performed when the specific frequencies of the targets are not known or are only partially known. Note that this form of searching, although described here as sequential, could also be simultaneous if a wideband, channelized receiver is used. Channelization in that case would normally also correspond to one channel width according to where in the spectrum the searching is taking place.

For directed search, specified frequencies are used for the searching. These frequencies form a list and the receiver tunes to each frequency in turn, looking for the presence of signals of interest. This mode is used when the frequencies are known.

Combinations of general and directed search are also possible. Indeed, this would likely be the normal mode of operation as even though many frequencies might be known ahead of time, some important targets may have changed frequency and those are not known. A typical example of this would be as a result of EA operations against a target net.

Both of these search strategies require the measurement of energy in the channel to which the receiver is tuned. Since noise is always present along with the signals, and since noise can only be described statistically, signal detection is a probabilistic process. The ability of a receiver to detect signals can only be described in statistical terms, even if the signals of interest are deterministic with parameters that are totally known.

Before any processing of a signal can occur, that signal must first be detected. That is it must first be ascertained if there is a signal present. More generally, detection refers to deciding which of two or more hypotheses is correct. Sometimes, however, signal detection is included as one of the outcomes of the signal classification problem, where the signal absent hypothesis is one of those tested. Any classification other than signal absent implies there is a signal present, signal detected.

There is another dimension to signal detection in communication EW system design, however. Since the target environments are typically noncooperative, it is often necessary to scan the RF spectrum looking for signals. When a signal is encountered, that signal is detected. Since the RF spectrum is everywhere channelized (although the size of the channels varies), this signal detection is normally accomplished by measuring the energy in these channels.

7.7.1.1 Hypothesis Testing

The presence of a signal $s(t)$ in a channel is characterized by two hypotheses, denoted by H_0 and H_1, depending on whether the signal is present or not:

$$\begin{aligned} H_0 &: r(t) = n(t) \\ H_1 &: r(t) = s(t) + n(t) \end{aligned} \tag{7.79}$$

In these equations, $r(t)$ is the received signal that consists of either $s(t)$ accompanied by noise $n(t)$, or just the noise alone. Hypothesis testing is the statistical process of estimating the validity of some hypothesis based on some set of measurements. Suppose that the measured parameters are manifest in a variable x. That is, the decision variable is $x = f(p_1, p_2, \dots, p_N)$ where p_i is a measured parameter.

When the decision rule is "decide the hypothesis that has the most likely probability of occurring," it is called the MAP criterion. Thus,

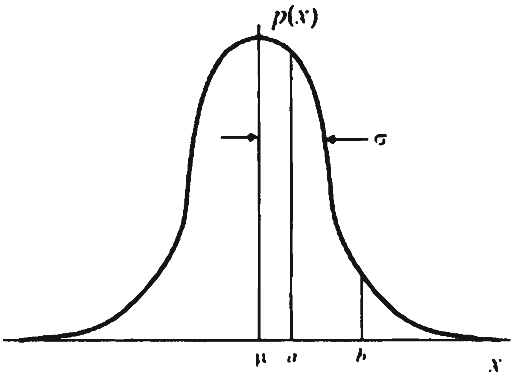


Figure 7.19 The probability of an event is given by the area under the density function between two points.

$$\begin{array}{lll} \text{Decide } H_0 \text{ if} & \Pr\{H_0|x\} > \Pr\{H_1|x\}, \text{ or } \Pr\{H_0|x\} \;/\; \Pr\{H_1|x\} > 1 \\ \text{Decide } H_1 \text{ if} & \Pr\{H_0|x\} < \Pr\{H_1|x\}, \text{ or } \Pr\{H_0|x\} \;/\; \Pr\{H_1|x\} < 1 \end{array} \tag{7.80}$$

Using the definition of conditional probabilities

$$\begin{aligned} &\Pr\{H_0|x\} = \Pr\{H_0\} \Pr\{x|H_0\} / \Pr\{x\} \text{ and} \\ &\Pr\{H_1|x\} = \Pr\{H_1\} \Pr\{x|H_1\} \;/\; \Pr\{x\} \end{aligned} \tag{7.81}$$

$\Pr\{H_0\}$ is the a priori probability of H_0 occurring at all, while $\Pr\{H_1\}$ is the a priori probability of H_1 occurring at all. Decide H_0 if

$$\frac{\Pr\{H_0|x\}}{\Pr\{H_1|x\}} = \frac{\Pr\{H_0\}\Pr\{x|H_0\}/\Pr\{x\}}{\Pr\{H_1\}\Pr\{x|H_1\}/\Pr\{x\}} > 1 \tag{7.82}$$

or

$$\frac{\Pr\{x|H_0\}}{\Pr\{x|H_1\}} > \frac{\Pr\{H_1\}}{\Pr\{H_0\}} \tag{7.83}$$

Recall from the definition of the probability density function (see Appendix A) that a probability is given by the area under the density function. That is, if $p(\xi)$ is the density function, then (referring to Figure7.19)

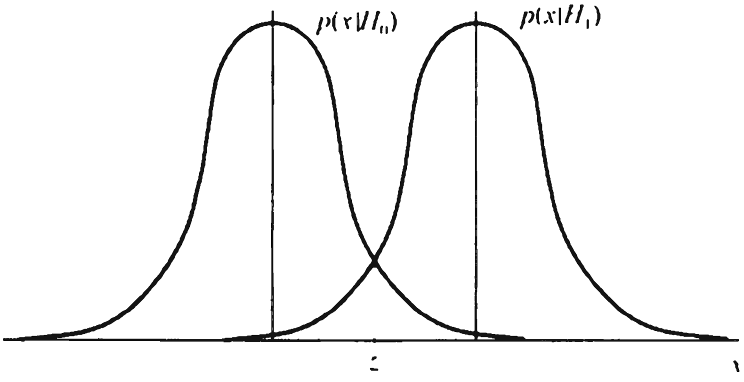


Figure 7.20 Probability density functions for H_0 and H_1.

$$\Pr\{a \le x \le b\} = \int_a^b p(\xi)d\xi \tag{7.84}$$

As $\Delta x = b - a$ becomes smaller and smaller, then

$$\lim_{\Delta x \to 0} \frac{\Pr\{x|H_0\}}{\Pr\{x|H_1\}} = \lim_{\Delta x \to 0} \frac{\int_a^b p(\xi|H_0)d\xi}{\int_a^b p(\xi|H_1)d\xi} = \frac{p(x|H_0)}{p(x|H_1)} \tag{7.85}$$

This ratio is known as the *likelihood ratio*, and is denoted by λ_0. Thus,

$$\lambda_1 = \frac{p(x|H_1)}{p(x|H_0)} > \frac{\Pr\{H_0\}}{\Pr\{H_1\}} \quad \text{for hypothesis } H_1 \tag{7.86}$$

$$\lambda_0 = \frac{p(x|H_0)}{p(x|H_1)} \ge \frac{\Pr\{H_1\}}{\Pr\{H_0\}} \quad \text{for hypothesis } H_0 \tag{7.87}$$

Although this analysis compares the ratios to a threshold value of one, it obviously can be trivially extended for any threshold ζ. This parameter then becomes a multiplier for the right side of these inequalities.

Suppose that $p(x \mid H_0)$ and $p(x \mid H_1)$ are as shown in Figure 7.20. Let $\Pr_m$ denote the probability of missed detection. This is the probability that, given that a signal is present (H_1), it is not detected. As shown in Figure 7.20, this is the area under the $p(x \mid H_1)$ curve to the left of the threshold point ζ. Thus,

$$\text{Pr}_\text{m} = \int_{-\infty}^{\zeta} p(\xi|H_1)d\xi \tag{7.88}$$

Let Pr_fa denote the probability of false alarm. This is the probability, given that there is no signal present (H_0), that the decision is made that there is. In Figure 7.20 this is the area under the $p(x\,|\,H_0)$ curve to the right of the threshold ζ. Thus,

$$\text{Pr}_\text{fa} = \int_{\zeta}^{\infty} p(\xi|H_0)d\xi \tag{7.89}$$

Both of these probabilities are a measure of an error being made. Thus, the overall probability of error, denoted by Pr_e, is the sum of these two quantities:

$$\text{Pr}_\text{e} = \text{Pr}_\text{m} + \text{Pr}_\text{fa} \tag{7.90}$$

The probability of detection, denoted here by Pr_d is equal to one minus the probability of missed detection, namely,

$$\text{Pr}_\text{d} = 1 - \text{Pr}_\text{m} = 1 - \int_{-\infty}^{\zeta} p(\xi|H_1)d\xi = \int_{\zeta}^{\infty} p(\xi|H_1)d\xi \tag{7.91}$$

7.7.1.2 Effects of Noise on Detection

The SNR is the dominating parameter to determine signal detectability. Suppose $p_\text{n}(v)$ represents the probability density function of the output power of a receiver when there is no signal present—that is, the output consists of noise only. Further suppose $p_\text{n+s}(v)$ represents the probability density function of this output when there is a signal present. Figure 7.21 shows the two receiver output density functions under these two assumptions. When there is a signal present, the mean value of the detector output will be larger than when there is no signal present. Here, the probability density function for noise only is assumed to have a zero average (mean) value.

Characteristics of receivers utilizing these concepts are typically presented on what are called ROC curves. Figure 7.22 shows a typical *receiver operating characteristic* (ROC) curve and clearly shows how Pr_d and Pr_fa can be traded off. If one wants to achieve a high Pr_d, then a high Pr_fa is typically required, and vice versa. If a low Pr_fa is desired, one usually must accept a low Pr_d. Optimizing both is usually not possible.

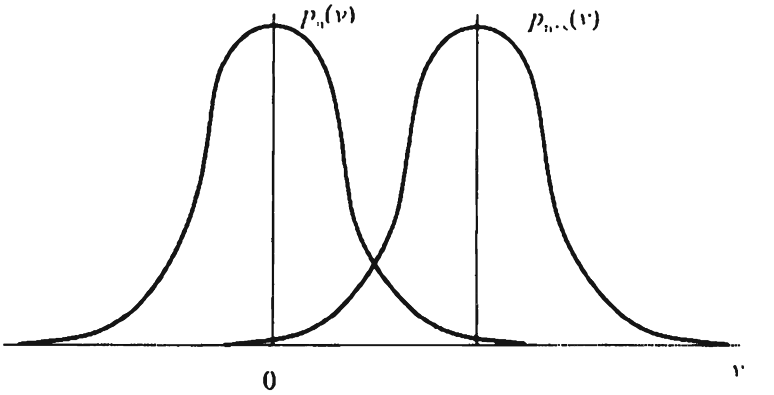


Figure 7.21 Detector output probability density functions with noise only [$p_n(\nu)$] and signal plus noise [$p_{n+s}(\nu)$].

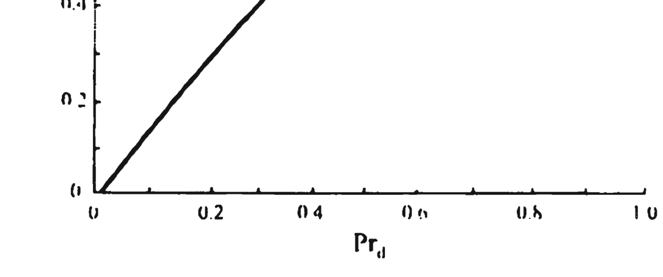


Figure 7.22 Typical ROC curve.

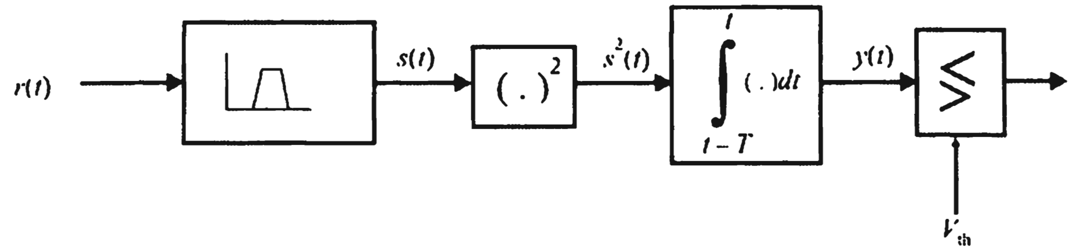


Figure 7.23 Block/flow diagram of a radiometer that measures energy over a time interval (0, *T*).

7.7.1.3 Radiometer

A diagram of a radiometer is shown in Figure 7.23 [19, 20]. Radiometers are frequently used as a standard against which other types of signal detectors are compared, although real implementations of radiometers are possible (or at least close approximations). A radiometer is an energy detector since the energy, E, in signal $s(t)$ is given by

$$E = \int_0^T s^2(t)\ dt \tag{7.92}$$

where it is assumed that the signal lasts only from 0 to T or, equivalently, the analysis of the signal is over the interval $(0, T)$. The radiometer is an optimum detector if it is assumed that the signal is a stationary process with a flat spectral density and the noise is Gaussian [21]. The input to the radiometer is the signal to be detected. It is sent through a bandpass filter, of width wide enough to pass $r(t)$ without significant distortion, yielding $s(t)$. This signal is then squared and integrated from 0 to T. After T seconds, the integrated results $y(t)$ are compared with a threshold V_{th}. If $y(t) > V_{\text{th}}$, then the signal is declared present. If $y(t) < V_{\text{th}}$, then the signal is declared absent.

For realistic communication EW systems, wideband radiometers are idealized devices, having limited practical use. What is more common is to measure the energy in a frequency channel using a narrowband radiometer. Such a configuration is often referred to as a channelized radiometer. The performance of such a configuration depends, along with other parameters, on how long the signal is present. This could be a considerable length of time for such signals as special mobile radios, for example, or it could be very short in the case of fast frequency-hopping targets.

The performance of a radiometer can be determined as follows [22]. Assuming there is noise only at the input, then the output statistics have chi-squared characteristics with $2TW$ degrees of freedom where T is the integration time and W is the noise bandwidth. If there is a signal present, then the output statistics are also chi-square but are noncentral. The noncentrality parameter is $2E/N_0$, where E is the energy in the input signal given by

$$E = P_s T \tag{7.93}$$

N_0 is the one-sided noise spectral density in W/Hz and P_s is the power in the signal.

When TW is large, the output statistics are approximately Gaussian from central limit theorem arguments. These statistics are characterized by a factor d, which is a measure of the output SNR. It is defined as

$$d = Q^{-1}(P_{fa}) - Q^{-1}(P_d) \tag{7.94}$$

where Q^{-1} is the inverse normal cumulative distribution function. For the radiometer the performance is given by

$$d^2 = \frac{T}{W}\left(\frac{P_s}{N_0}\right)^2 \tag{7.95}$$

Thus, the SNR required to achieve the specified Pr_{fa} and Pr_d is given by

$$\frac{P_s}{N_0} = d\sqrt{\frac{W}{T}} \tag{7.96}$$

The parameter d can be determined numerically. Some values for typical performance parameters of interest are shown in Figure 7.24.

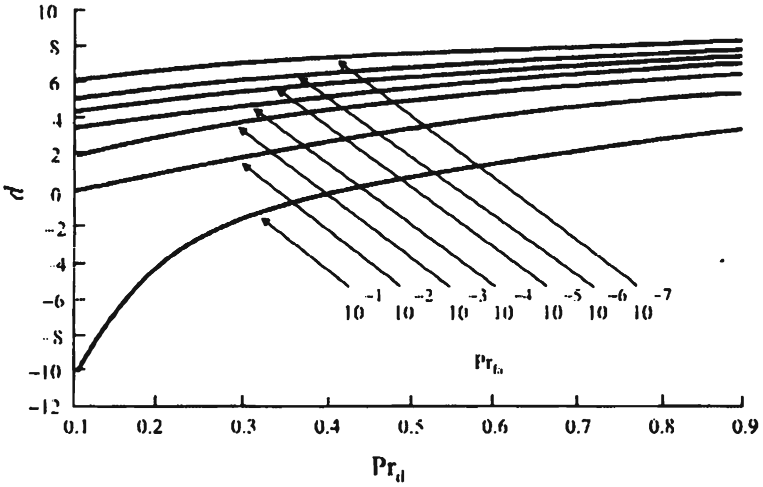


Figure 7.24 A measure of the output SNR of a radiometer. (*Source:* [22].)

Typically for communication EW applications, *TW* is not large. Some numbers for the low VHF frequency range might be $T = 1$ ms and $W = 25$ kHz for channelized radiometers. Thus, $TW = 25$. In this case the above analysis generates values for *d* that are too optimistic [22]. This is optimistic from an intercept point of view—from a communicator's point of view, they are pessimistic. In that case, a factor is provided in Woodring to multiply *d* to correct the analysis. Charts of this factor, η, for two cases, $\text{Pr}_\text{d} = 0.9$ and $\text{Pr}_\text{d} = 0.1$, are shown in Figures 7.25 and 7.26. Therefore, for small *TW*, ηd is used instead.

7.7.1.4 Detection of Frequency-Hopping Signals

The wideband radiometer discussed above is an optimum detector for signals if there is little known about the signals. A channelized radiometer is constructed if the input is first filtered. One approach for constructing a detector for frequency-hopping signals is to assemble *G* bandpass radiometers that operate on the incoming signal, where *G* is the number of channels used by the frequency-hopping target. Each of these channels has a bandwidth matched to the channelization of the target radio. After N_h hop intervals of duration T_h each, the outputs of the channelized radiometers are compared with a threshold η, and under hypothesis H_0 (signal absent), the output should be less than η; and under H_1 (signal present), it should be greater than η. If more is known about the signal of interest, then the radiometer is not an optimum detector for the signal. The optimum detector will, in general, be different for each different type of signal. These optimum detectors can become quite complex. Figure 7.27 shows the configuration for the optimum detector for slow frequency-hopping signals, which employ multiple frequencies within a hop (*multiple FSK*, MFSK) and also employ continuous phase within a hop [23]. In this case, the detector is optimumin the *average likelihood ratio* (ALR) sense. For this figure, E_h is the energy in each hop, $N_0/2$ is the two-sided noise level, $I_0(\ )$ is the modified Bessel function of the 0th kind, and N_d is the number of data symbols in each hop. The frequency corresponding to the *i*th channel is denoted by f_i while θ_i corresponds to the local oscillator phase offset. The function $\varphi(d_i, t)$ is the baseband data modulation.

Thus, for each of the *G* channels, N_d noncoherent detectors are implemented. In each of these N_d detectors, there is a noncoherent detector required for each possible data sequence that can occur. Each of these detectors is configured as shown in Figure 7.27. Each consists of two parallel channels of multipliers, integrators and squaring circuits. These two channels are added and then the square root of the result is computed. This is followed by multiplication by the SNR estimate, which is followed by computation of the Bessel function. All of these computations are executed in channel as shown. Within each channel the outputs of the data pattern channels are added together, followed by addition of the results from each channel. After the N_h hops, the product of the results from each hop is computed and compared with the

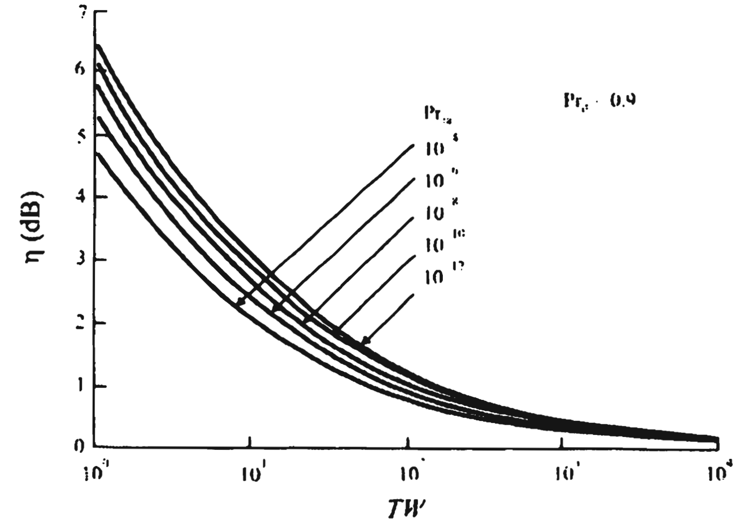


Figure 7.25 Adjustment factor η for *d* when $Pr_d = 0.1$. (*Source:* [22].)

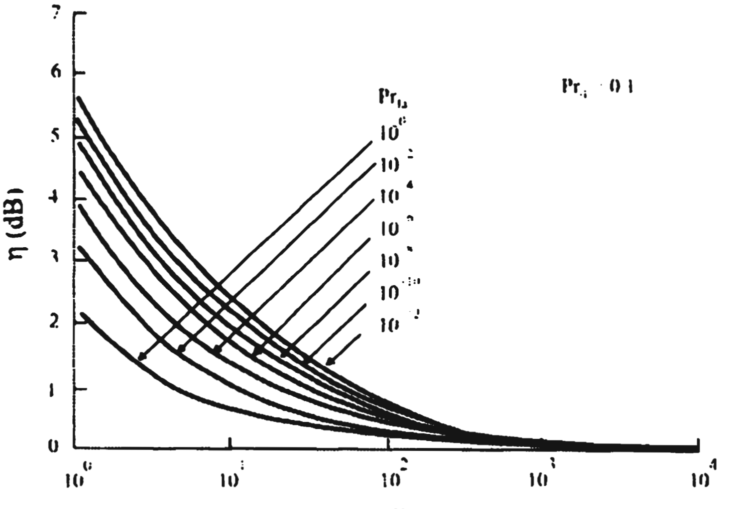


Figure 7.26 Adjustment factor η for small values of *TW* when $Pr_d = 0.9$. (*Source:* [22].)

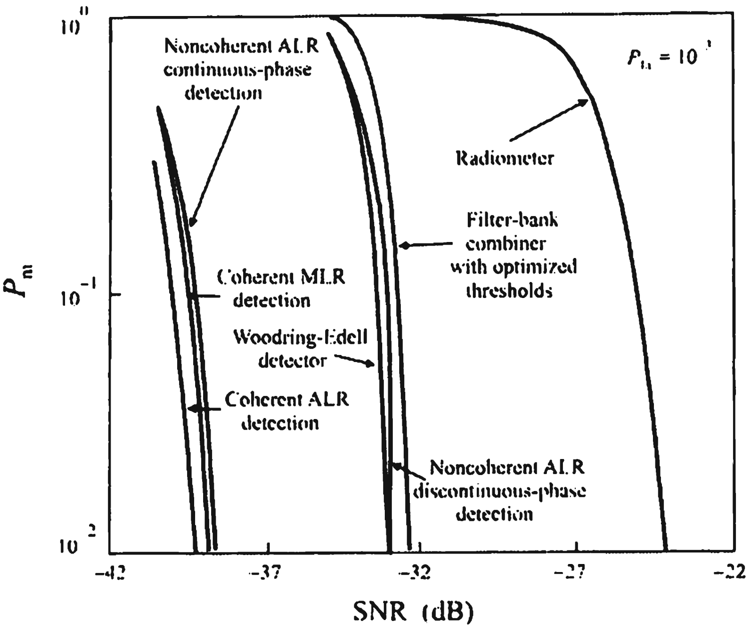


Figure 7.28 Probability of missing a detection for the detector shown in Figure 7.19 compared with other forms of slow frequency-hopping signal detectors. (*Source:* [25]. © IEEE 1994. Reprinted with permission.)

parallel in each possible input sequences whereas for the latter, the most likely threshold.The difference between optimization based on the ALR and the *maximum likelihood ratio* (MLR) is that for the former, the average ratio is determined for all sequence is determined. MLR detection is suboptimal compared with ALR, becausenot all of the possible input sequences are considered in the analysis. Performance of the above optimum detector compared with other techniques is illustrated in Figure 7.28 [24].

The above detector is optimum for slow frequency-hopping targets with continuous phase from one symbol to the next within a hop, but noncoherent from hop to hop. The case of fast frequency-hopping targets was analyzed by Beaulieu et al. [25]. The most common definition of slow frequency hopping versus fast is based on the number of data bits per hop. If there is more than one such data bit per hop, it is called slow frequency-hopping, while if there is more than one hop per data bit (less than one data bit per hop), it is called fast frequency-hopping. The optimum detector structure for this case is shown in Figure 7.29 [23].

7.7.2 DSSS Interception Detection

In this section we compare the detection performance of three energy detectors that are routinely used to detect the presence of DSSS signals. The radiometer is such a detector that does not depend on knowing the parameters of the *signal of interest* (SOI). The other two detector structures, the chip rate detector and the frequency

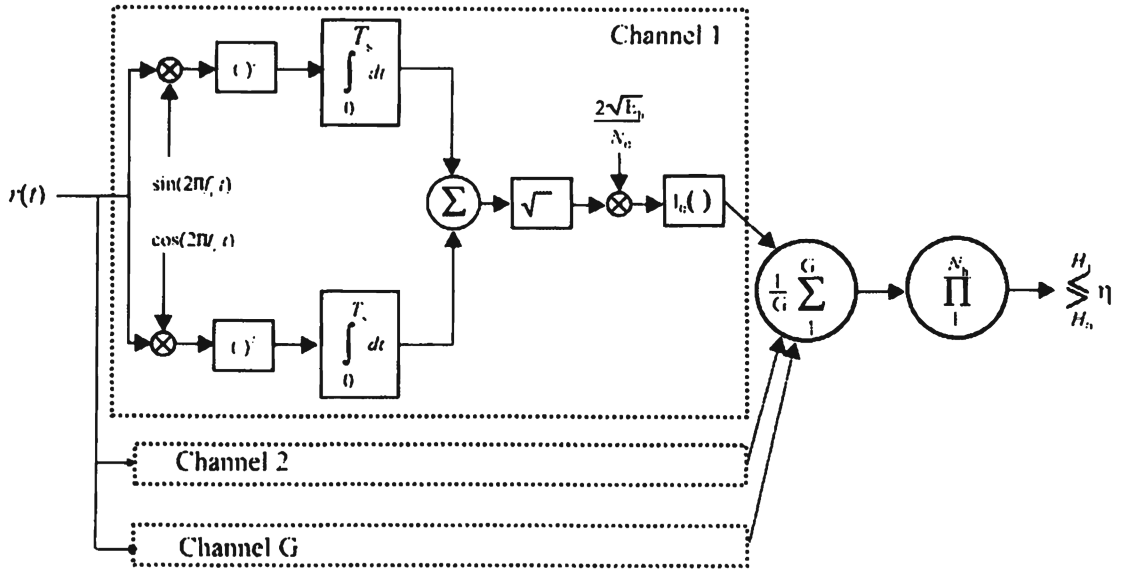


Figure 7.29 Architecture of an optimum detector for fast frequency-hopped signals. (*Source:* [23]. © IEEE 1992. Reprinted with permission.)

doubler, do depend on parameters of the SOI, as we will show in the sequel. These parameters are rather easily extracted, however.

The radiometer measures the energy in the input signal directly by squaring the signal and integrating for a symbol time T_S. This is compared to a threshold β that is a function of the desired false alarm probability, P_{fa}. The chip rate detector multiplies the signal with the complex conjugate of a delayed version of the signal. The delay is the duration of one chip. The net result out of the multiplier is a deterministic tone and noise. The frequency doubler feature generator multiplies the signal with a delayed version of the signal. The result out of the multiplier is a deterministic tone at twice the chip rate and noise. The Neyman-Pearson criterion is frequently used for signal detection. It determines the probability of detection, P_d, as a function of input SNR at a specified probability of false alarm, P_{fa}. When identical observed data are input to two intercept detectors, the best detector is determined by the one with the largest P_d.

While the Neyman-Pearson criterion closely represents typical system requirements, it is often difficult to evaluate, and obscures important effects in tradeoff analyses. Consequently other performance measures, such as output SNR is often used instead. The best of many detectors, or the best realization of a particular detector can be determined by the one that maximizes the output SNR.

We show below that while the output SNR is useful for evaluating and optimizing a particular detector class, it may be inappropriate when applied to different detector classes. This is because the distribution of the decision variable may differ for different detector classes.

Rifkin published an analysis of the intercept behavior of three types of detectors for signal intercept [26]. In particular he examined the radiometer, illustrated in Figure 7.30, and two forms of feature detectors for detection DSSS BPSK/QPSK signals. The two-features detectors are similar in structure as shown in Figure 7.31. The feature generator for the frequency doubler is illustrated in Figure 7.32 while the feature generator for the chip rate detector is shown in Figure 7.33. In this section we examine these two performance measures, output SNR and probability of detection, for these three intercept detectors.

We make the following assumptions:

- The noise is AWGN with known, or estimated, power;
- The power spectral density of the signal of interest is known;
- Typical for DSSS, the input SNR is much less than unity;
- The observation time is much greater than the chip duration.

As determined by Rifkin, specific examples are identified in which the radiometer's performance, using the Neyman-Pearson criterion, is inferior to that of the feature detectors operating at equal output SNR for the decision variable. When the Neyman-Pearson criterion is used with equal *input* SNRs, however, the radiometer always performs better than the feature detectors.

7.7.2.1 Detector Structures

Figures 7.30 and 7.31 show the structures for all three detectors. The input to each detector is a BPSK-modulated signal in AWGN noise with one-sided power spectral density N_0. Higher order PSK (e.g., QPSK) could also be used as the DSSS modulation and the results do not change. The BPSK-modulated chips are assumed to be rectangular, with their phase determined from a pseudo-noise sequence that is clocked at a rate much higher than the symbol rate. The input SNR for each detector, p_i, is given by $p_i = E_c/N_0$, where E_c is the energy per chip, and satisfies $p_i << 1$. Each intercept detector includes a complex multiply, which generates signal × signal, signal × noise, and noise × noise terms. The assumption of low input SNR dictates that both the signal × signal and signal × noise terms be ignored.

The integrated complex multiply output from the radiometer can immediately be compared with a threshold to make a detection decision. However, the tones generated by the feature detectors must be further processed to form the decision variable as seen in Figures 7.32 and 7.33. Quadratic detection forms the magnitude squared of its input, and is useful when the phase of the tone is unknown, which is normally the case for noncooperative interception. The magnitude-squaring operation substantially changes the pdf, however.

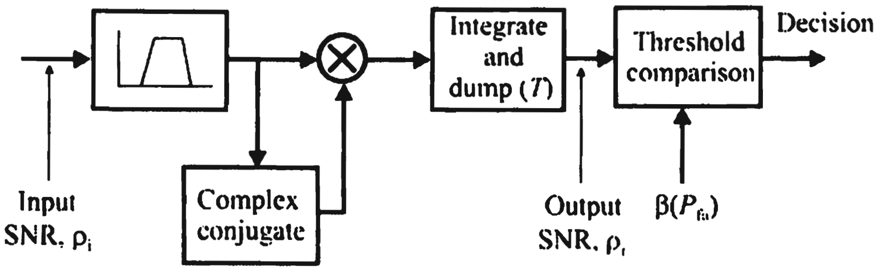


Figure 7.30 Rifkin radiometer. (*Source:* [26]. © IEEE 1994. Reprinted with permission.)

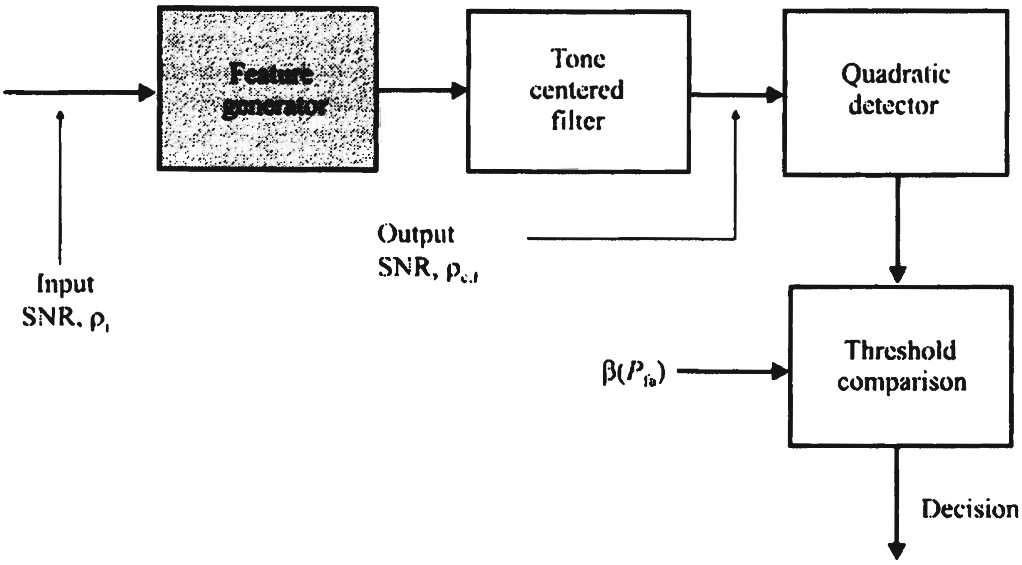


Figure 7.31 Feature detectors. (*Source:* [26]. © IEEE 1994. Reprinted with permission.)

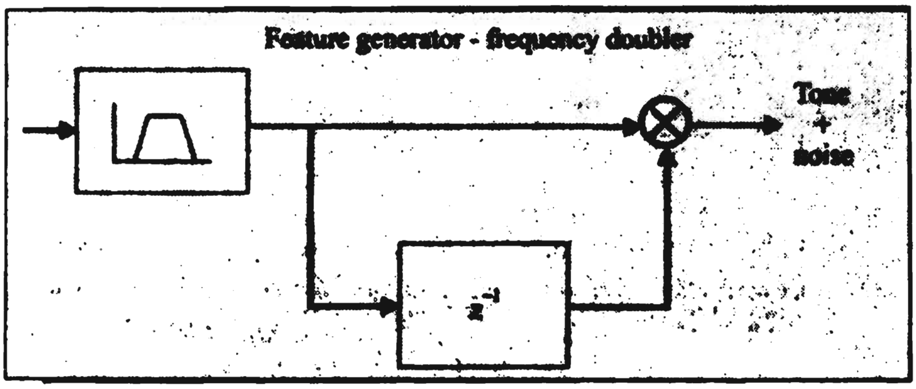


Figure 7.32 Feature generator for the frequency doubler. (*Source:* [26]. © IEEE 1994. Reprinted with permission.)

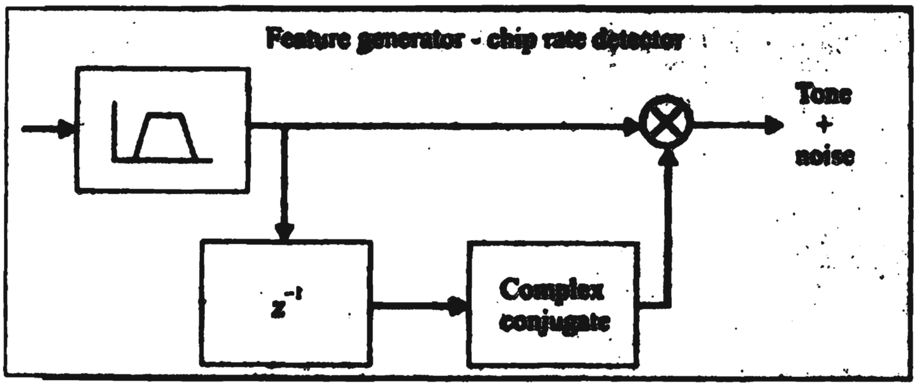


Figure 7.33 Feature generator for the chip rate detector. (*Source:* [26]. © IEEE 1994. Reprinted with permission.)

The relationship between the input SNR and the output SNR for each detector are given by

$$\rho_r = 0.667 N_c \rho_i^2, \qquad \text{radiometer} \qquad (7.97)$$

$$\rho_c = 0.101 N_c \rho_i^2, \qquad \text{chip rate detector} \qquad (7.98)$$

$$\rho_f = 0.333 N_c \rho_i^2, \qquad \text{frequency doubler} \qquad (7.99)$$

For the radiometer, if the signal-absent distribution out of the integrating filter is approximated by the normal distribution $\mathcal{N}(\mu_n,\sigma_n)$ with mean μ_n and variance σ_n^2, and the equivalent signal present distribution is represented by $\mathcal{N}(\mu_S,\sigma_S)$, then the threshold β, needed to obtain a particular P_{fa} is given by [26]

$$\beta(P_{fa}) = \sqrt{2}\sigma_n \operatorname{erfc}^{-1}(2P_{fa}) + \mu_n \qquad (7.100)$$

where erfc(.) is the complementary error function.

The probability of detection is given by:

$$P_d = \frac{1}{2}\operatorname{erfc}\left\{\frac{\sigma_n}{\sigma_S}\left[\operatorname{erfc}^{-1}(2P_{fa}) - \sqrt{\frac{\rho_r}{2}}\right]\right\} \qquad (7.101)$$

where the output SNR of the decision variable for the radiometer, ρ_r, has been defined as:

$$\rho_r = \frac{(\mu_S - \mu_n)^2}{\sigma_n^2} \qquad (7.102)$$

Since the input SNR << 0, both the signal × signal and signal × noise components at the output of the complex multiplier are negligible, and hence $\sigma_n \approx \sigma_S$. Thus, the probability of detection as a function of the output SNR, (7.101), simplifies to

$$P_d = \frac{1}{2}\operatorname{erfc}\left[\operatorname{erfc}^{-1}(2P_{fa}) - \sqrt{\frac{\rho_r}{2}}\right] \qquad (7.103)$$

For the frequency doubler and chip rate detector, the output of the complex multiplier is a deterministic tone and noise. Since the SNR << 0, the signal × noi and signal × signal terms can be ignored and by central limit theorem arguments, assuming $N_c >> 1$, the pdf of the signal at the multiplier is approximated as Gaussian with a nonzero mean due to the tone. If the detection decision variable is formed by squaring and adding the two independent r.v.s the result is a Ricean distribution. This leads to the threshold for these two detectors of

$$\beta = -2\sigma^2 \ln(P_{fa}) \tag{7.104}$$

where σ^2 is the variance of the decision variable. The probability of detection is then:

$$P_d = \int_{-\ln(P_{fa})}^{\infty} e^{-(z+\rho_{c,f})} I_0\left(2\sqrt{z\rho_{c,f}}\right) dz \tag{7.105}$$

The deterministic components out of the mixer that results from the signal of interest differ substantially for the three detectors. For the radiometer this component is a constant (centered about 0 Hz). For the chip rate detector this component is a series of tones at multiples of the chip rate, and for the frequency doubler it is also a series of tones centered at twice the carrier frequency, and offsets at multiples of the chip rate. THC chip rate detector and the frequency doubler are generally classified as feature detectors, since they both use signal-specific information (i.e., the keying rate and the carrier frequency) of the signal of interest. As shown in Figures 7.32 and 7.33, the feature detectors use complex multiplication of the input with a delayed version of itself to generate the feature (i.e., one or more tones) that can be detected and then compared with a threshold.

7.7.2.2 Results

Figure 7.34 shows the relationship between the input and output SNRs for the three detectors considered when $N_c = 10^6$. As is clear from Figure 7.34, the radiometer shows the best detection performance for all input SNRs. However, the different distributions for the decision variable leads to other conclusions. Figure 7.35 illustrates P_d as a function of the output SNR. The curves cross and therefore sometimes the radiometer is better while sometimes the feature detectors are better.

This is further substantiated with Figure 7.36. In this case, the probability of detection is plotted versus the *output* SNR, ρ_r or $\rho_{c,f}$. The curves cross or not depending on the value of P_{fa} chosen, indicating that sometimes the radiometer is better and in other circumstances the feature detectors are better. As remarked above,

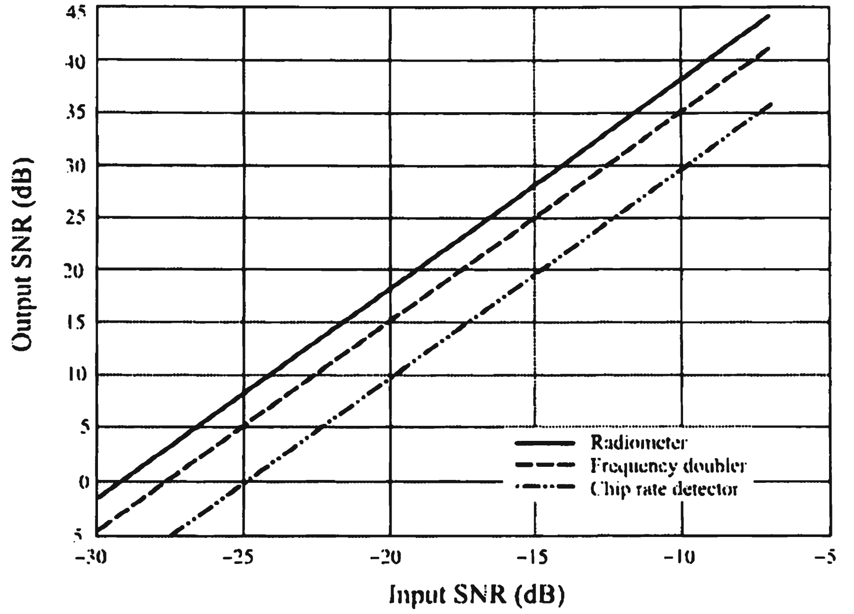


Figure 7.34 Input output SNR. ($N^c = 10^6$) (*Source:* [26]. © IEEE 1994. Reprinted with permission.)

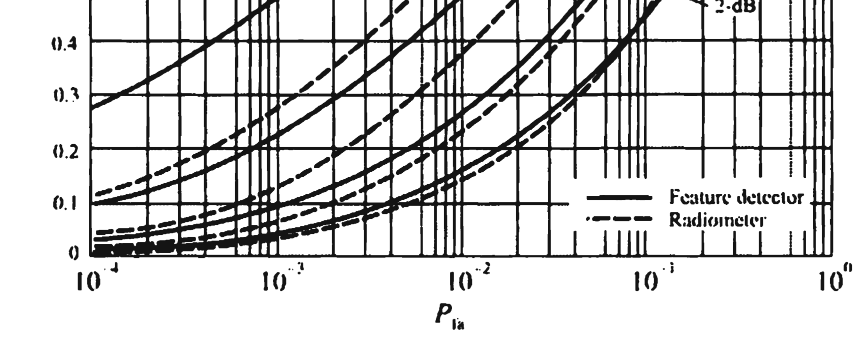


Figure 7.35 Rifkin ROC. (*Source:* [26]. © IEEE 1994. Reprinted with permission.)

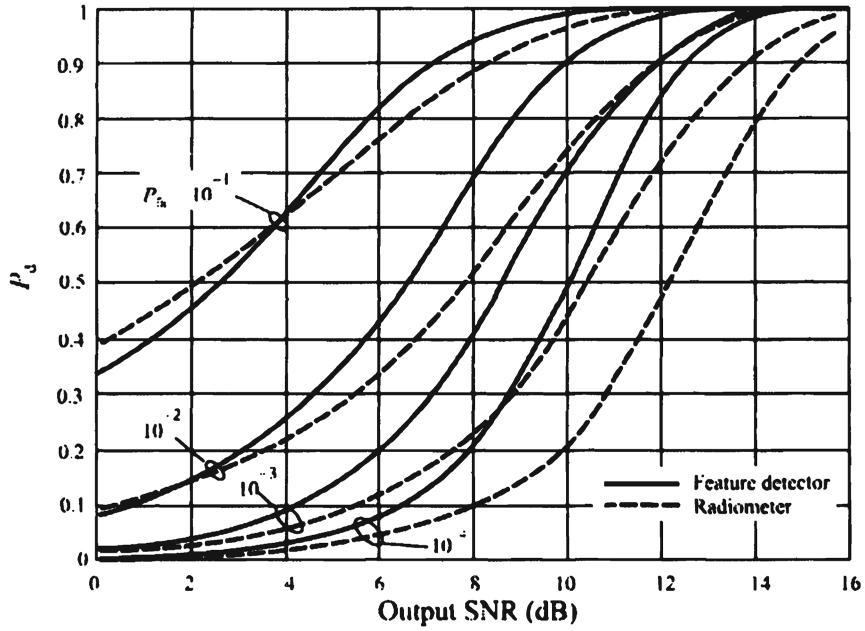


Figure 7.36 Probability of detection, P_d, versus SNR. (*Source:* [26]. © IEEE 1994. Reprinted with permission.)

however, when comparing the probability of detection against the *input* SNR, the radiometer is always better (the radiometer is an optimal detector in AWGN).

7.7.3 Signal Classification

Another important function that is performed in the receiving subsystem of communication EW systems is signal classification. Classifying signals can help in determining the EOB by indicating what type of communication equipment is in use. Knowing the type of equipment an adversary has sometimes will also indicate who that adversary is, or perhaps indicate the measure of lethality or danger it presents to friendly forces.

Signal classification can be further broken down into several types of processes as listed here (not necessarily mutually exclusive or exhaustive):

- Modulation recognition (e.g., AM, FM, or PM);
- Signal type recognition (e.g., analog or digital);
- If digital, then which type (e.g., BPSK or QPSK);
- If modem, then which standard (e.g., V.32 or V.90);
- If multiplexed, then which type (e.g., FDM or *pulse position modulation*, PPM);
- If FDM, the type of signal in each channel;

- Specific emitter identification (e.g., to the serial number of the transmitter).

In general, there are two approaches to address the signal classification problem. The first approach is somewhat ad hoc. Signal features are determined largely by reasoning about what might make sense—amplitude histograms, for example, might make a good discriminate for AM versus constant modulus signal types. The first step in these processes is called feature extraction. The computed features that are extracted in the first step are then provided to processors that perform pattern recognition that attempts to discriminate one signal type from another. Neural networks, being devices that do pattern recognition reasonably well and fast, are often applied to this form of signal classification.

The second general approach is based on probability theoretic notions, also referred to as decision theoretic. Likelihood ratios are determined and hypotheses are established based on probability density functions of random signals. Cost functions are determined and the minimum costs are achieved by the likelihood ratios. This makes these approaches optimum in that sense; costs are minimized.

Experimentation by simulation has shown that the latter of these approaches offers more of a possibility of operation at lower SNRs than the former does. The former approaches tend to fall off in performance when the SNR falls below about 15 dB or so. The latter has, at least in simulations, successfully operated at SNRs less than zero. Low-SNR operation of such classifiers is important when long target ranges are involved. The more robust a classifier is, the more applicability it will have.

7.7.3.1 Pattern Recognition Approaches

When discussing signal detection, the notion of hypothesis testing was introduced in the context of a binary hypothesis problem: determining the presence of a signal in a channel. The signal classification problem can be cast in similar terms, although generalization to multiple hypotheses is required. In that case, there are N hypotheses: selecting one of $N-1$ signal types and selecting the Nth choice, which is "none of the above."

As a simple example, suppose that the classes of interest are the following: AM DSB, AM SSB, wideband FM ($BW > 100$ kHz), and narrowband FM ($BW < 100$ kHz). Then the hypotheses are as follows.

- H_1: AM DSB;
- H_2: AM SSB;
- H_3: Wideband FM;
- H_4: Narrowband FM;

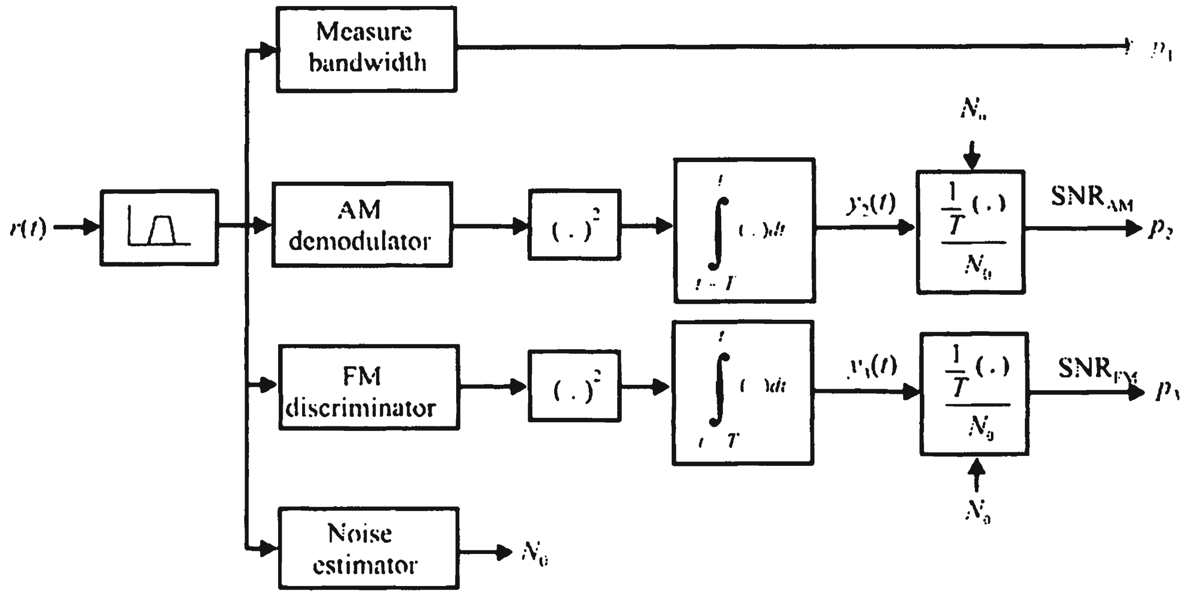


Figure 7.37 Simple pattern recognition signal classifier.

- H_5: Other (cannot determine or something other than the specified signal types).

The first step in these approaches is to extract relevant features. For this simple example, suppose that these features consist of the following

- p_1 = Bandwidth of the predemodulated signal;
- p_2 = SNR of the output of an AM demodulator;
- p_3 = SNR of the output of an FM discriminator.

This is shown diagrammatically in Figure 7.37. It is tacitly assumed here that the width of the bandpass filter is adequate to pass the signal relatively undistorted. In this case, BW = 150 kHz might be about right.

The logic for selecting among the hypotheses using these features might be as follows. In this simple scheme, analog AM SSB would be very narrowband, so the bandwidth of the signal would be an important discriminator. Narrowband FM would have a narrower measure of bandwidth than wideband FM as well. These, coupled with the SNR at the output of the AM and FM demodulators, would determine the type of signal present. In those cases where there is potential confusion, the classifier indicates indeterminate results, which is probably the desired output in such circumstances. The threshold level for the SNR calculations in this example is 6 dB. In practice, this is a threshold that would be determined from experimentation. The resultant logic that implements this thought process is presented in Table 7.3.

Table 7.3 Hypotheses Selections for the Example Classifier

p_1	p_2	p_3	**Choose**
< 5 kHz	< 6 dB	< 6 dB	H_6
< 5 kHz	< 6 dB	> 6 dB	H_4
< 5 kHz	> 6 dB	< 6 dB	H_2
< 5 kHz	> 6 dB	> 6 dB	H_5
> 5 kHz and < 100 kHz	< 6 dB	< 6 dB	H_6
> 5 kHz and < 100 kHz	< 6 dB	> 6 dB	H_4
> 5 kHz and < 100 kHz	> 6 dB	< 6 dB	H_1
> 5 kHz and < 100 kHz	> 6 dB	> 6 dB	H_5
> 100 kHz	< 6 dB	< 6 dB	H_6
> 100 kHz	< 6 dB	> 6 dB	H_3
> 100 kHz	> 6 dB	< 6 dB	H_1
> 100 kHz	> 6 dB	> 6 dB	H_5

The nemesis to accurate signal classification is the same as other communication signal processing problems—noise. This is manifest in the above example by the presence of N_0, the estimate of the noise floor used in the calculation of the SNR. This noise was discussed in Chapter 2. Since in all real instances of communication signals, noise is always present, the usual parameters calculated for signal classification are statistical. The various statistical moments of the signal over a short time interval, for example, are common parameters, with the most common being second order statistics but sometimes higher order are used.

Nandi-Azzouz Classifier

Nandi and Azzouz have proposed an algorithm for the automatic classification of communication signals [27–29]. In their case, the problem was to distinguish among 13 analog and digital modulation types: AM, FM, multiple (= 2, 4) FSK, multiple (= 2, 4) ASK, DSB, LSB, USB, multiple (= 2, 4) PSK, vestigial sideband (VSB), and combined modulations. Combined in this case refers to a signal that has both substantial amplitude and phase modulations present. Nine parameters are used to perform the signal classification and these parameters are based on measurements of the signal in the time domain as well as the frequency domain. The flow diagram for this classifier is shown in Figure 7.38 while the parameter definitions are given in Table 7.4. In most cases, the data from which these parameters are obtained is normalized to some relevant value.

Each of these parameters was compared with threshold values that were derived by simulation. The thresholds are indicated in Table 7.5, where at each stage a parameter is compared to its threshold, thus dividing the classes into two groups at

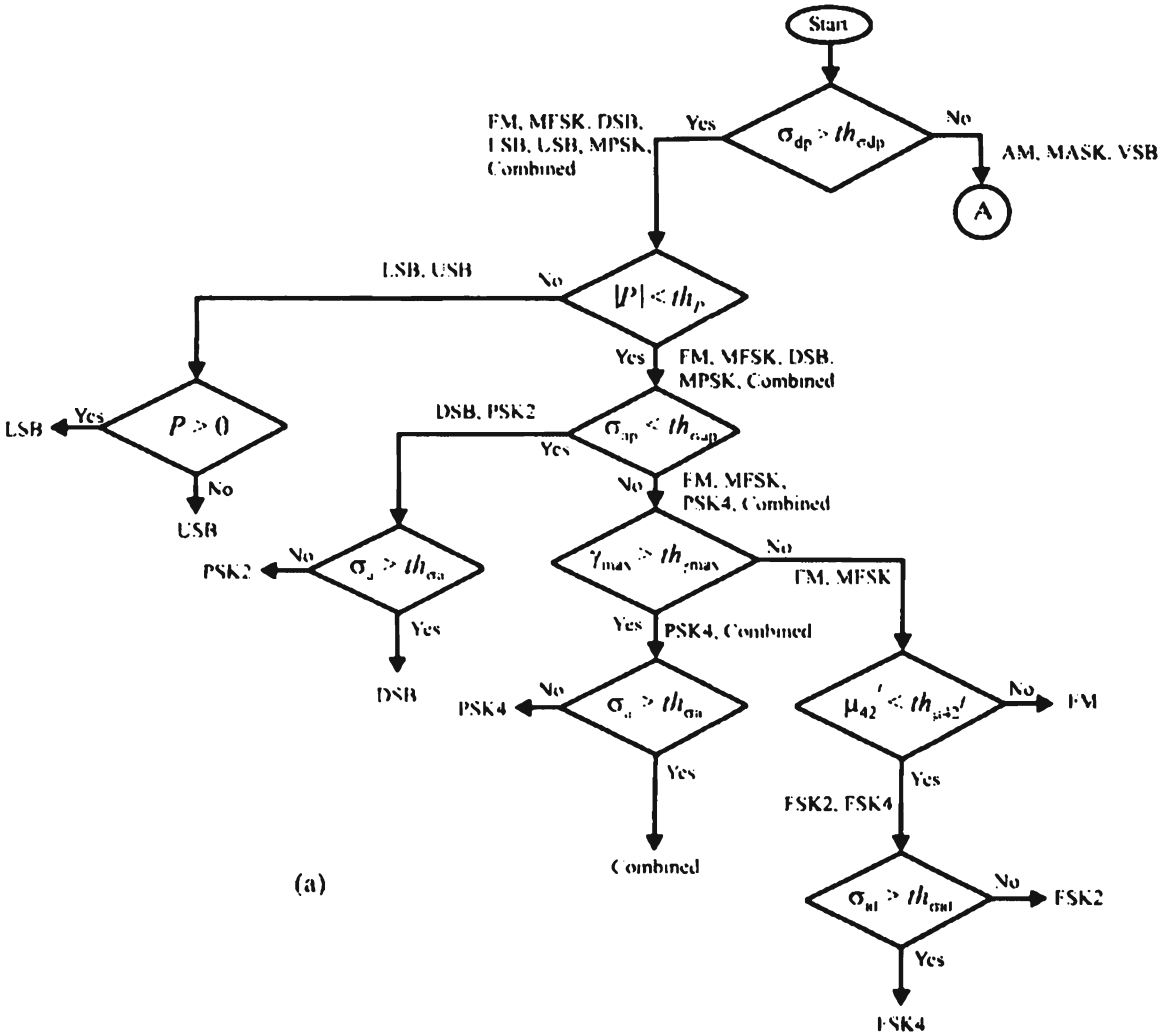

Start
$\sigma_{dp} > th_{\sigma dp}$
Yes
No
FM, MFSK, DSB, LSB, USB, MPSK, Combined
AM, MASK, VSB
A
$|P| < th_p$
No
LSB, USB
Yes
FM, MFSK, DSB, MPSK, Combined
$P > 0$
Yes
LSB
No
USB
$\sigma_{ap} < th_{\sigma ap}$
DSB, PSK2
Yes
No
FM, MFSK, PSK4, Combined
$\sigma_a > th_{\sigma a}$
No
PSK2
Yes
DSB
$\gamma_{max} > th_{\gamma max}$
No
FM, MFSK
Yes
PSK4, Combined
$\sigma_a > th_{\sigma a}$
No
PSK4
Yes
Combined
$\mu_{42}' < th_{\mu 42}'$
No
FM
Yes
FSK2, FSK4
$\sigma_{af} > th_{\sigma af}$
No
FSK2
Yes
FSK4
(a)

Figure 7.38 (a) Flow diagram for the Nandi-Azzouz signal classifier; and (b) remainder of that flow diagram.

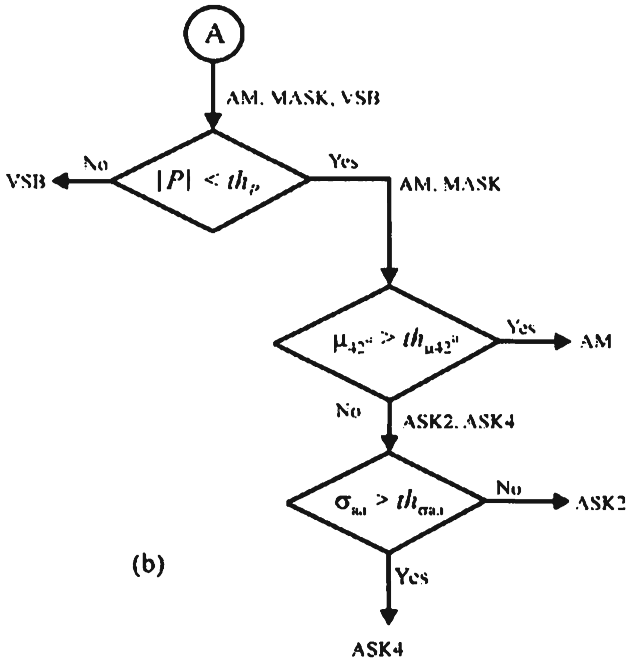


Figure 7.38 (Continued.)

Table 7.4 Parameters Used for the Nandi-Azzouz Pattern Recognition Signal Classifier

Parameter	Definition
γ_{max}	Maximum of the PSD of the normalized-centered instantaneous amplitude
σ_{ap}	Standard deviation of the absolute value of the instantaneous phase
σ_{dp}	Standard deviation of the instantaneous phase
P	Symmetry of the spectrum of the signal
σ_{aa}	Standard deviation of the absolute value of instantaneous amplitude
σ_{af}	Standard deviation of the absolute value of the instantaneous frequency
σ_a	Standard deviation of the instantaneous amplitude
μ_{42}^a	Kurtosis of the instantaneous amplitude
μ_{42}^f	Kurtosis of the instantaneous frequency

Table 7.5 Threshold values of the parameters for the Nandi – Azzouz Classifier.

Parameter	Threshold Value (th$_i$)
γ_{max}	2.5
σ_{ap}	$\pi/5.5$
σ_{dp}	$\pi/6$
P	0.6
σ_{aa}	0.13
σ_{af}	2.15
σ_a	2.03
μ_{42}^a	0.25
μ_{42}^f	0.4

each such stage, gradually separating each modulation until only a single type remains in the group. Results of this classifier against the target signals set are reported only for correct signal classification in [27], except for the digital modulations. These results are repeated in Figure 7.39. These parameters yielded very good performance tracked with a PLL that extracts the operating frequency over time, determining the moving average of frequency.

As a last note on the performance of this classifier, the SNRs considered are typical of those obtained from an airborne EW system, but are much too high to consider for ground applications. In fact, 15–20 dB SNRs are not unusual for even standoff airborne systems because signals propagate much better when one (or both) of the antennas is significantly elevated. For ground-to-ground applications, however, the power decreases substantially more rapidly and SNRs of 15–20 dB are difficult to produce.

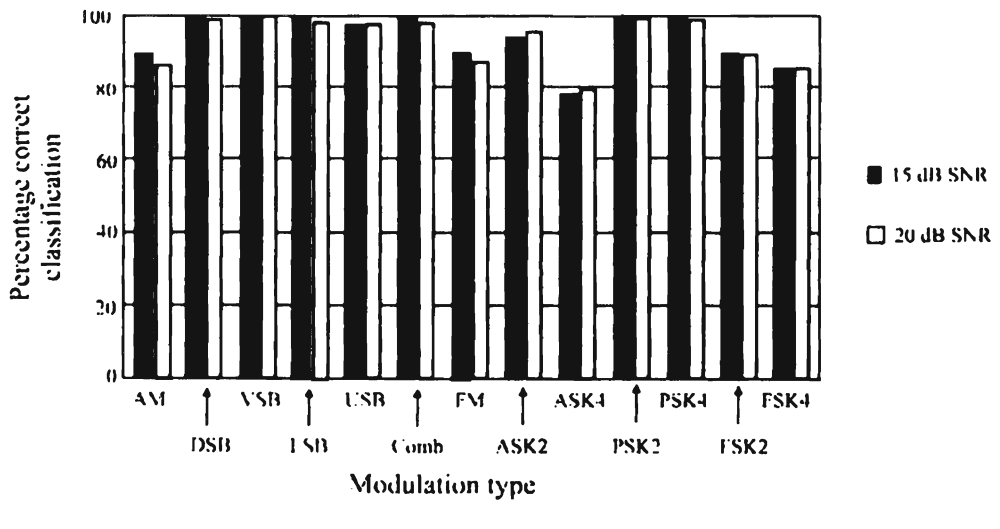


Figure 7.39 Performance of the Nandi-Azzouz classifier. (*Source:* [20], © 1998 IEEE. Reprinted with permission.)

Table 7.6 Values of Phase Shift According to the Type of Modulation

Modulation Type	$\phi(k)$
CW	0
BPSK	0, π
QPSK	0, $\pm\pi/2$, π
BFSK	$\pm 2\pi f_d k$
QFSK	$\pm\pi f_d k$, $\pm 2\pi f_d k$

Assaleh-Farrell-Mammone Classifier

A technique was devised by Assaleh et al. [30] that classifies digitally modulated signals. It is based on calculating the autocorrelation coefficients of the predetected data. The *intermediate frequency* (IF, predemodulation, also referred to as predetected) signal is represented by

$$x(k) = s(k) + n(k) \tag{7.106}$$

where $s(k)$ are samples of the signal to be identified and $n(k)$ are samples of noise. The signal $s(k)$ is given by

$$s(k) = A\cos\left(2\pi f_c k + \phi(k) + \theta\right) \tag{7.107}$$

where f_c is the signal frequency, θ is the phase, and the parameter ϕ varies according to the signal type as given in Table 7.6.

The autocorrelation estimates of this signal are given by

$$\hat{R}_{rr}(k) = \sum_{n=0}^{M} r(n)r(n+k) \tag{7.108}$$

These autocorrelation estimates are the coefficients of a polynomial of dimension N given by

$$1 - a_1 z^{-1} - a_2 z^{-2} - \cdots - a_N z^{-N} \tag{7.109}$$

The average instantaneous frequency is related to the roots of this polynomial, z_i, by

$$\hat{R}_{rr}(k)=\sum_{n=0}^{M} r(n)r(n+k) \tag{7.110}$$

$$f_i=\frac{f_s}{2\pi}\tan^{-1}\left(\frac{\mathrm{Im}(z_i)}{\mathrm{Re}(z_i)}\right) \tag{7.111}$$

and the bandwidth of this root is

$$BW_i=-\frac{f_s}{\pi}10\log_{10}\left[\frac{1}{\mathrm{Im}^2(z_i)+\mathrm{Re}^2(z_i)}\right] \tag{7.112}$$

In these expressions, f_s is the sample rate. It is these two parameters that the algorithm uses to classify the signals shown in Table 7.6. The classification algorithm is shown in Figure 7.40. The authors report that usually a second-order ($N = 2$) model is all that is necessary in this algorithm.

The simulated performance of this classifier against the signals considered is shown in Figure 7.41. For these results, the SNR was 15 dB, which is relatively high for many EW applications (it is typical to have the requirement to prosecute distant targets and 10 dB SNR or less is normal). Almost perfect classification ensued under the conditions of the experiment. The algorithm was not tested, however, for robustness in noisy environments.

Whelchel-McNeill-Hughes-Loos Classifier

Implementation of a modulation classifier using neural networks was described by Whelchel et al. [31]. The signals of concern for their classifier were AM, AM SC, FM, CW, WGN, and QPSK. WGN is *wideband Gaussian noise*. They describe two experiments in which the SNR of the signals were 23 dB and 30 dB, representing relatively high values of SNR.

As with other pattern classification approaches, the first step is to extract the features from the incoming predetection signal. The features used in this classifier are based on several statistical moments and are given in Table 7.7. Two types of neural networks were trained: a back propagation network and a counterpropagation network [31]. Another variation that was tried was to connect (or not) the input layer of the neural network to the output layer. The performance results are shown in Figure 7.42 for the former network configuration to illustrate the type of performance achieved. These results are similar to the performance of the other pattern recognition approaches. The advantage of the neural network, however, is thespeed.

Neural networks are massively parallel processors and compute results very rapidly. They emulate one model of how the human brain functions, with processing

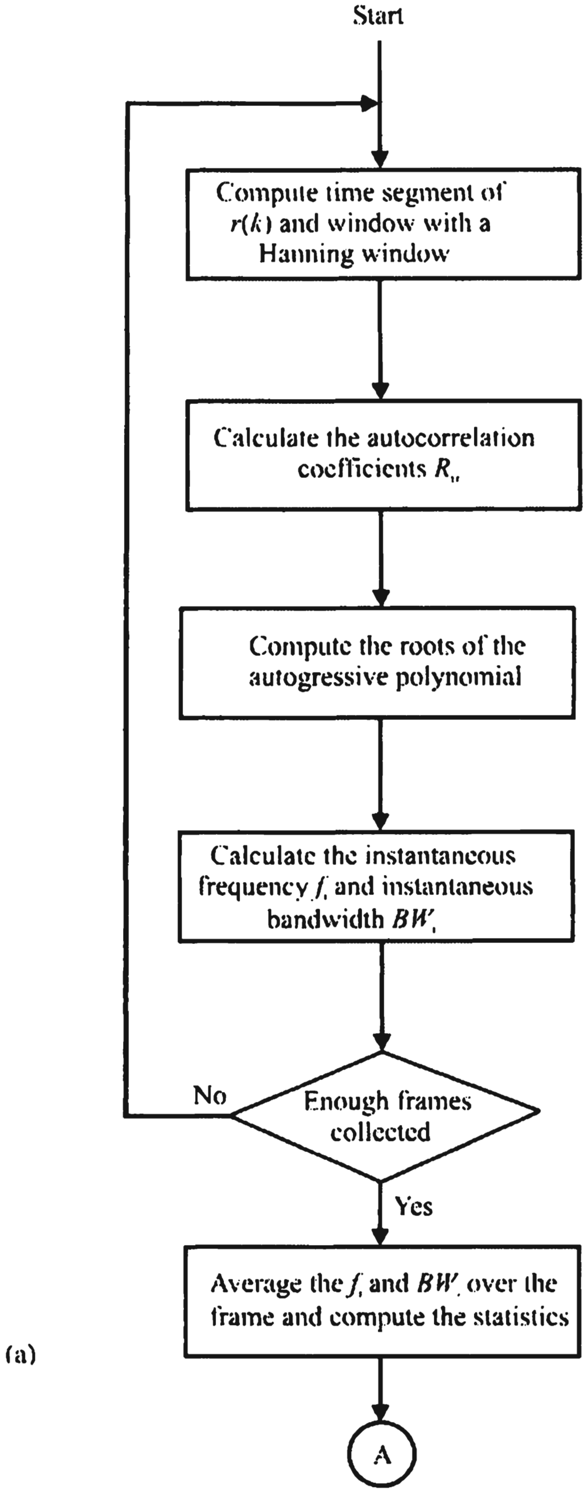

Start
Compute time segment of $r(k)$ and window with a Hanning window
Calculate the autocorrelation coefficients R_n
Compute the roots of the autogressive polynomial
Calculate the instantaneous frequency f_i and instantaneous bandwidth BW_i
Enough frames collected
No
Yes
Average the f_i and BW_i over the frame and compute the statistics
(a)
A

Figure 7.40 (a) Flow diagram for the Assaleh-Farrell-Mammone signal classifier; and (b) remainder of that classifier. (*Source:* [30]. © IEEE 1992. Reprinted with permission.)

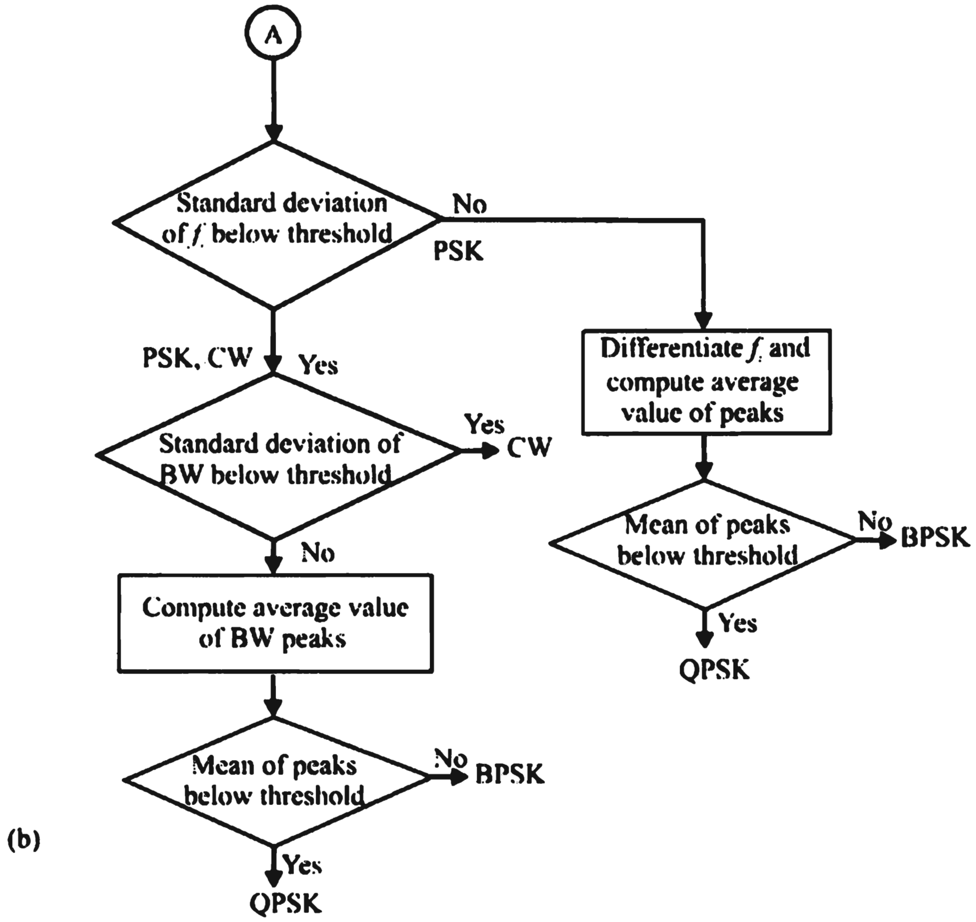

A
Standard deviation of f_i below threshold
No
PSK
PSK, CW
Yes
Standard deviation of BW below threshold
Yes
CW
No
Compute average value of BW peaks
Mean of peaks below threshold
No
BPSK
Yes
QPSK
Differentiate f_i and compute average value of peaks
Mean of peaks below threshold
No
BPSK
Yes
QPSK
(b)

Figure 7.40 (Continued.)

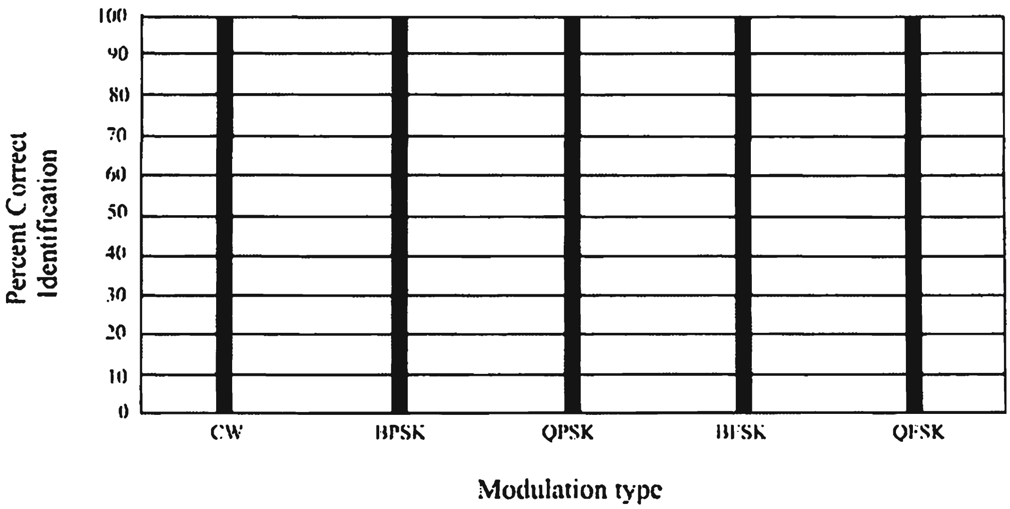


Figure 7.41 Simulated performance results of the Assaleh-Farrell-Mammone classifier. (*Source:* [30]. © IEEE 1992. Reprinted with permission.)

Table 7.7 Statistical Parameters Used in the Whelchel et al. Neural Network Signal Classifier

Amplitude variance
Amplitude skew
Amplitude kurtosis
Phase variance
Phase skew
Phase kurtosis
Frequency variance
Frequency skew
Frequency kurtosis

Source: [30].

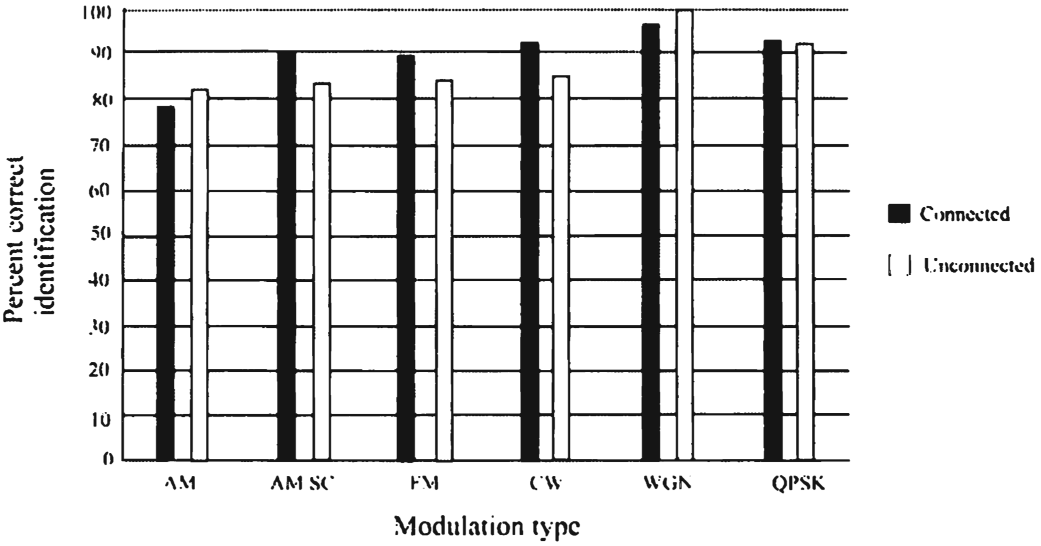


Figure 7.42 Performance of the Whelchel et al. neural network signal classifier. (*Source:* [31]. © IEEE 1989. Reprinted with permission.)

nodes that are interconnected with synapses. Many interconnects per node are implemented which facilitate the rapid computations. Neural networks are probably best known for their performance at pattern recognition [32].

7.7.3.2 Decision Theoretic Approaches

The decision theoretic signal classifiers establish likelihood ratios as discussed in Section 7.6.1.1 and compute the most likely signal type present by choosing the signal corresponding to the largest likelihood ratio. Since the likelihood ratios are based on statistical properties of the signal, they produce optimal decisions, with optimality depending on the particular ratio used.

Kim-Polydoros Classifier

Kim and Polydoros describe a decision-theoretic approach to discriminating between BPSK and QPSK [33]. The *quasi-likelihood ratio* (qLLR), which they determined for this discrimination, using their notation, is given by

$$\text{qLLR} = \left(\Sigma_{\text{I}} - \Sigma_{\text{Q}}\right)^2 + 4\Sigma_{\text{IQ}}^2 \tag{7.113}$$

where

$$\Sigma_{\rm I} = \sum_{n=1}^{N} r_{{\rm I},n}^2 \quad \Sigma_{\rm Q} = \sum_{n=1}^{N} r_{{\rm Q},n}^2 \quad \Sigma_{\rm IQ} = \sum_{n=1}^{N} r_{{\rm I},n} r_{{\rm Q},n} \tag{7.114}$$

when the complex envelope of the output of a matched filter with input $r(t)$ is given by

$$\tilde{r}_n = r_{{\rm I},n} + j r_{{\rm Q},n} \tag{7.115}$$

and the components given by

$$r_{{\rm I},n} = \int_{(n-1)T_s}^{nT_s} r(t)\cos(2\pi f_c t)dt \; r_{{\rm Q},n} = \int_{(n-1)T_s}^{nT_s} r(t)\sin(2\pi f_c t)dt \tag{7.116}$$

$n = 1, 2, \ldots, N$ and N is the number of symbols processed, while f_c is the frequency of the carrier.

The authors of [28] compared the performance of this classifier with two others. The *square law classifier* (SLC) to which they refer simply squares the input signal. Since $\cos^2(x) = ½[1 + \cos(2x)]$, this effectively doubles the frequency of the signal. If a significant component is found at this double frequency, then it is known that the input frequency was at x. For BPSK, the double frequency term shows up at twice the input data rate, while for QPSK it is at four times the input data rate. The other type of simple classifier discussed was the *phase-based classifier* (PBC), which operates directly on the phase changes from symbol to symbol. A histogram is generated of these phase changes and decisions about whether the signal is BPSK or QPSK are based on this histogram. The results of this approach are shown in Figure 7.43. These are simulated results with the number of symbols included, $N = 100$. As seen, better than 80% correct classification is possible even with the SNR as low as –5 dB with the qLLR algorithm. That algorithm significantly outperforms the other two at the low SNRs considered here.

Sills Classifier

Sills compared the performance of coherent versus noncoherent maximum likelihood modulation classification when the modulations of interest were three PSK and three QAM signals [30]. In particular, the modulations investigated are shown in Figure 7.44. The coherent maximum-likelihood classifiers were the traditional types that maximize the probability of a particular modulation type given the output of a coherent detector. For the coherent case, it was assumed that all the parameters of the

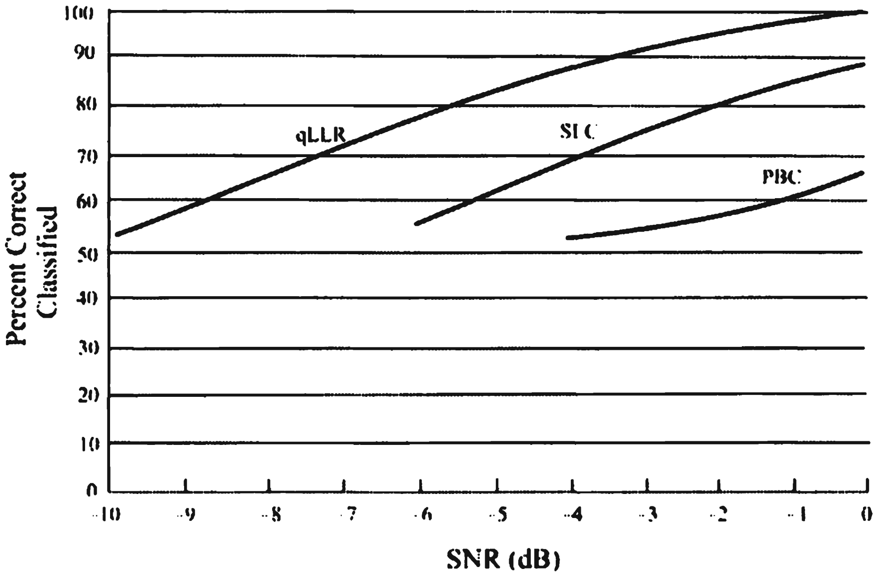


Figure 7.43 Performance of the Kim and Polydoros classifier. (*Source:* [33]. © IEEE 1988. Reprinted with permission.)

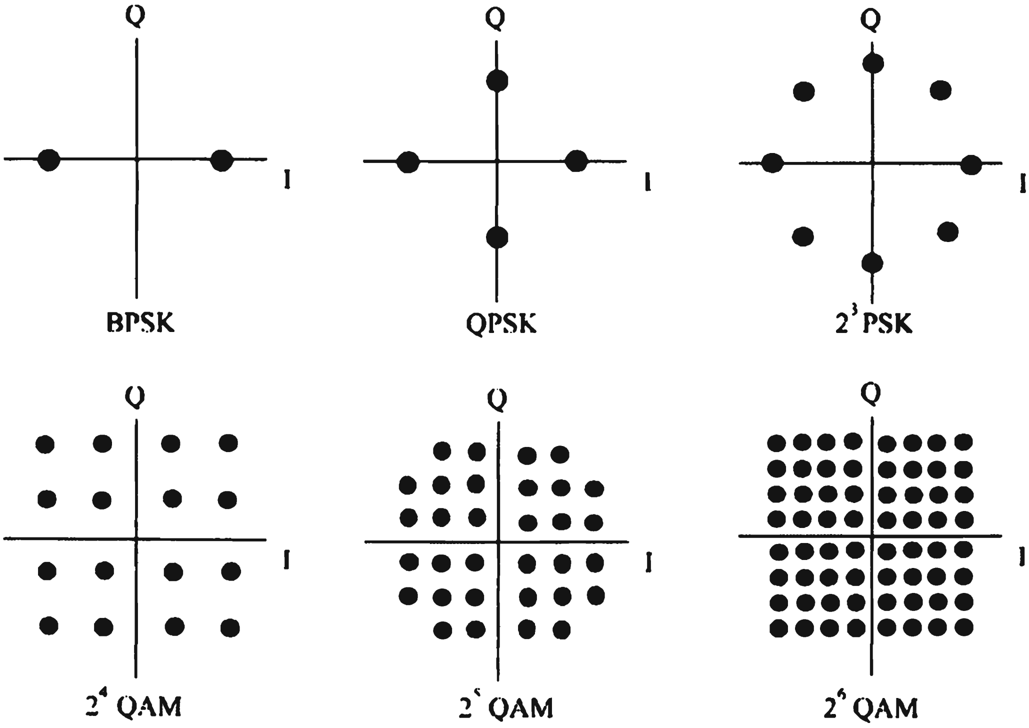


Figure 7.44 Constellations corresponding to standard modulations considered by Sills for signal classification. (*Source:* [34]. © IEEE 1999. Reprinted with permission.)

signal were known, including the carrier phase. The results of the simulation are shown in Figure 7.45. All of the signals were classified correctly at least 90% of the time whenever the SNR was greater than 10 dB.

As aptly pointed out, however, measuring the carrier phase for QAM is very difficult at these signal levels. Because of this, a second simulation was performed where it was assumed that the carrier phase was not known. Again, maximum-likelihood detection was assumed. These results are shown in Figure 7.46. There is about a 3 dB loss in performance from the coherent case, a number not unlike that associated with other communication problems when comparing coherent versus incoherent detection

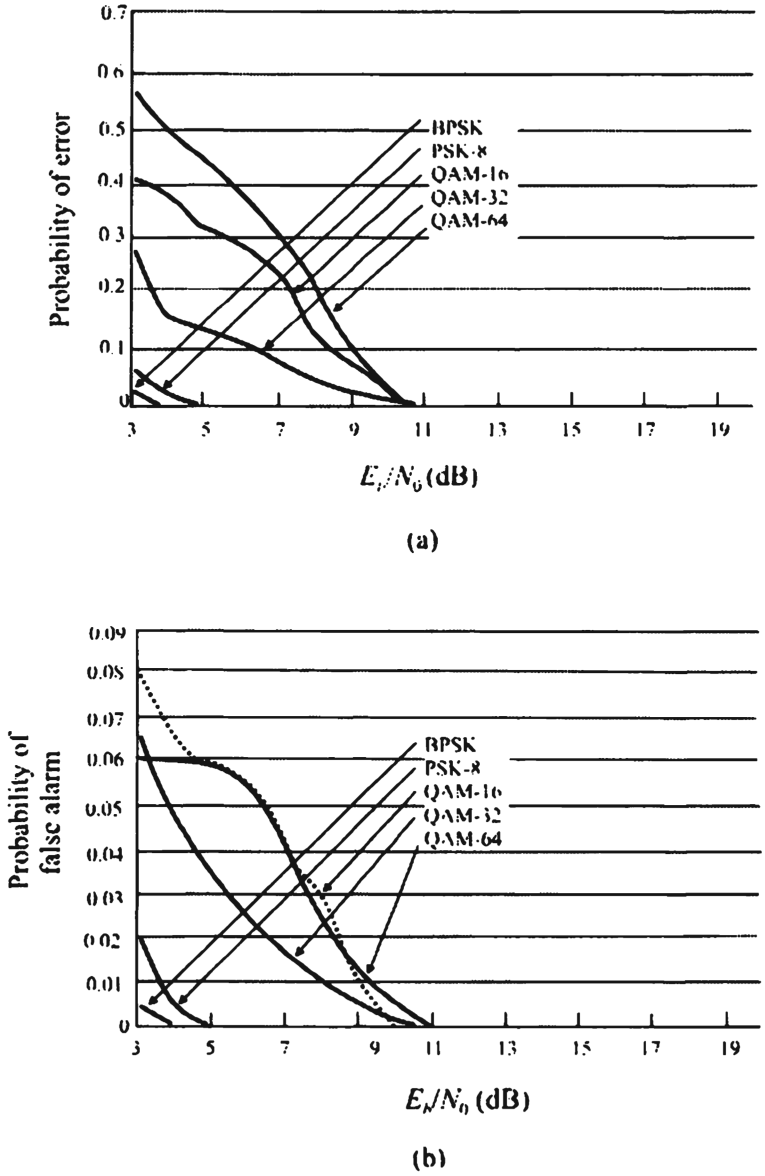
7.7.4 Recognition/Identification

Several functions in communication EW systems can be automated, at least to some degree. Those functions that can recognize or identify entities are prime candidates for such processing; however, automating the processes can be difficult. Usually the entity sets involved are limited in size; otherwise, performance degrades. The general flow diagram for recognition processing is shown in Figure 7.47. The input to these processes, in general, is different depending on the type of processing to be performed. In the case of voice recognition, for example, the input would be speech.

In the flow diagram shown in Figure 7.47, features are first extracted from the input signal. Feature extraction is the process of determining from the input signal those parameters that can be used to discern between different entities and, conversely, are consistent with the same entity. These features are compared with the features stored in the local database. In one scenario (open identification) if there is a sufficiently close match of the new features to a database entry, then a successful match is declared; otherwise, the new features are added to the database. In another scenario (closed identification), if there is a match, then success is declared, but if there is no match, then the conclusion is simply that the entity is not in the database—that is, the new features are not added to the database.

A flow diagram for the verification process is illustrated in Figure 7.48. It is somewhat different from the identification process in that only one comparison is required—the features are compared with the specific entity to be verified.

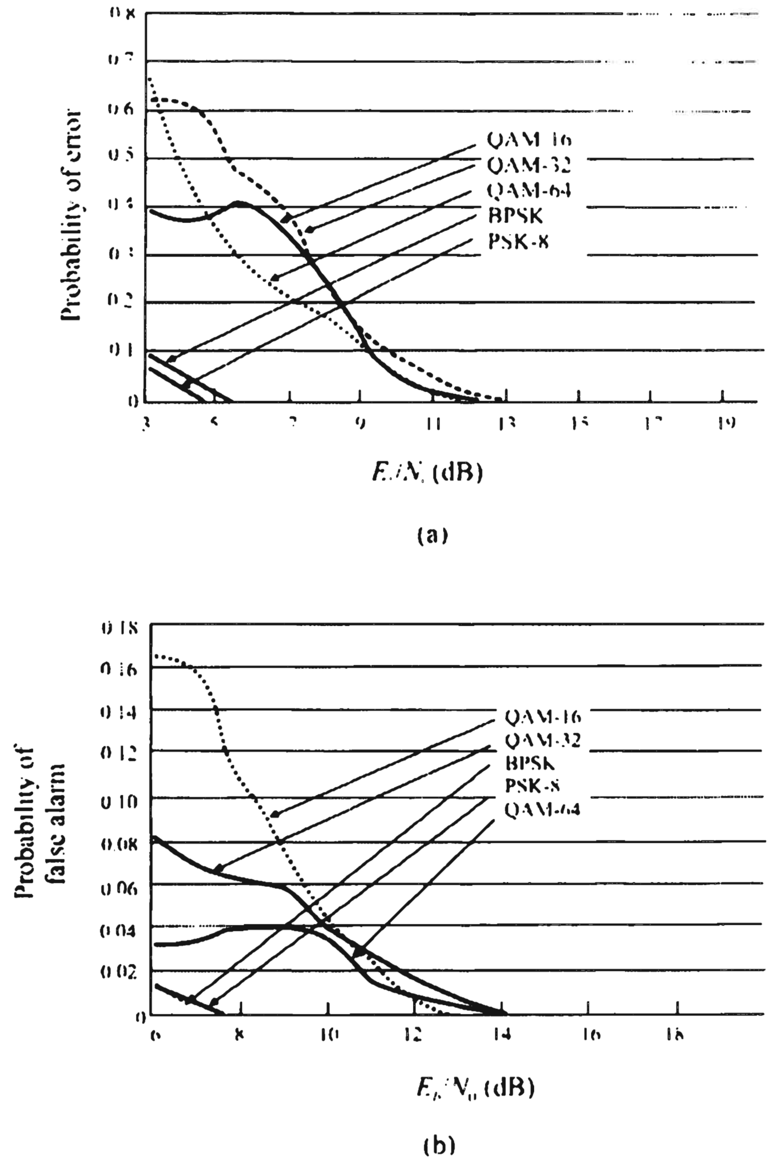
Open identification is when the set of entities under consideration is not limited. Some form of identification would then be associated with all new entities encountered. Closed identification is when the set of entities is limited, and the question to be answered is whether the current entity belongs to the set.



(a)

(b)

Figure 7.45 Sills maximum-likelihood classification results when coherent detection was assumed: (a) the probability of correct classification and (b) the probability of false alarm. (*Source:* [34]. © IEEE 1999. Reprinted with permission.)



(a)

(b)

Figure 7.46 Sills noncoherent maximum-likelihood classification. Shown in (a) are the probabilities of correct classification while (b) shows the probabilities of false alarm. (*Source:* [34]. © IEEE 1999 Reprinted with permission.)

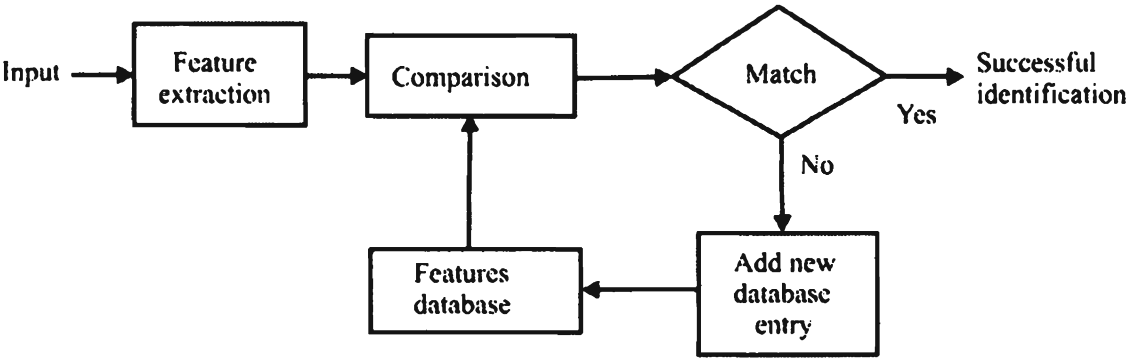


Figure 7.47 Flow diagram of the recognition/identification process.

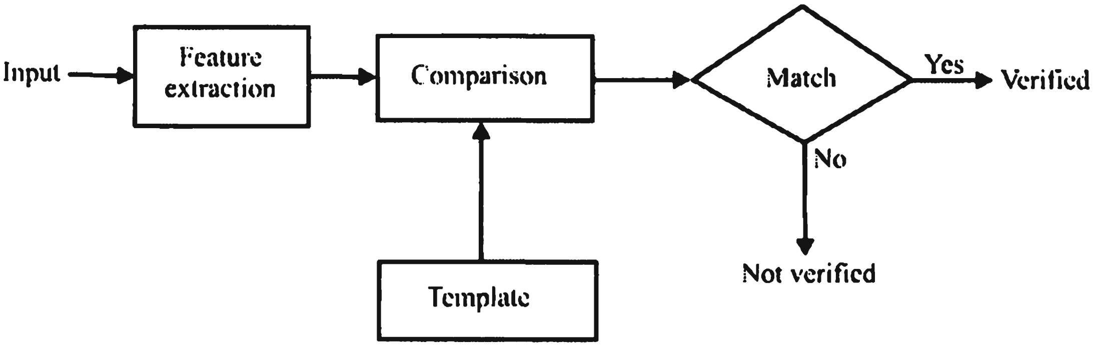


Figure 7.48 Flow diagram of the verification process.

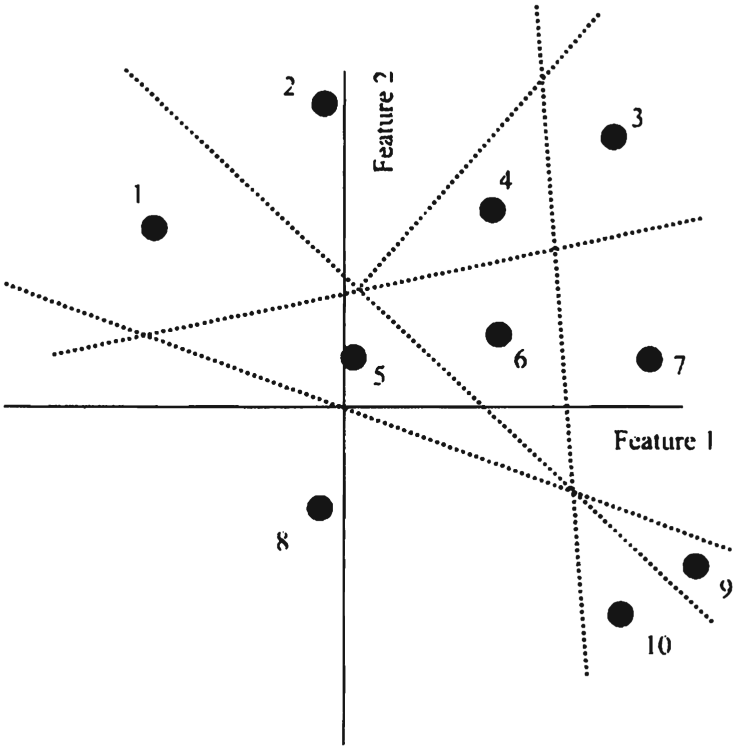


Figure 7.49 Segmenting a feature space for identification. In this example there are two features that make up the feature space and the regions are linearly separable.

7.7.4.1 Mathematical Modeling

Just as in other processing where entities are divided into groups based on features, the features form a space that must be segregated into regions. Each such region corresponds to a single decision. For just two features the feature space is illustrated in Figure 7.49. The dark circles represent the mean value of the two features in each region and the boundaries divide it into separate regions. Each feature must be sufficiently robust in order to be consistent with the same entity and sufficiently different from other entities. Each of the regions shown corresponds to a combination of features associated with each individual entity. There could be gaps between regions where undecided is the conclusion and it is desired to indicate this indecision. Neural networks and fuzzy logic can be successfully applied to the recognition problem. In reality the space would be N-dimensional and the features would form a hyperspace.

Statistical Modeling

The distance between an arbitrary point x in feature space and a point μ representing a specific entity can be computed in several ways. Perhaps the most common is the

Euclidean distance given by

$$d_{\mathrm{E}}^2 = (x - \mu)^2 \tag{7.117}$$

An alternative is the Mahalonobis distance of $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$ given by

$$d_{\mathrm{M}}^2 = (\bar{x} - \mu)^T \Sigma^{-1} (\bar{x} - \mu) \tag{7.118}$$

where μ and Σ are the mean and covariance of the training features, respectively, and $\bar{x}$ is the average of the feature vectors.

The points of equi-Euclidean distance from a point $(x_1, x_2, ..., x_n)$ form a hypercircle of radius d_{E} ,with the point at its center. The points equi-Mahalonobis distant from this point form a hyperellipsoid. The axes are determined by the eigenvectors and eigenvalues of the covariance matrix Σ. There is an advantage of using the Mahalonobis distance over the Euclidean distance in some problems. Which to use depends on how the features are distributed in the feature space.

Probabilistic Modeling

An alternative way to use the features to determine a match is to use probability densities. If p_i are the continuous probability densities and $p_i(x)$ is the likelihood that a feature x is generated by the ith entity, then

$$\Pr\{\text{identity} = i | x\} = \frac{p_i(x)}{p(x)} \Pr_i \tag{7.119}$$

where $\Pr_i$ is the probability that the input came from the ith entity and $p(x)$ is the probability of the feature x occurring from any entity. This is Bayes' rule, whose major shortcoming, as mentioned, is that the a priori probabilities must be known.

7.7.4.2 Source Recognition

This function attempts to identify the source of the information being transmitted. This source could be, for example, a person speaking, in the case of speaker identification, a person operating a Morse code key, or the computer providing digital data through a modem. Each component in a communication system has its own characteristics; however, they are not always distinguishable. The closer to the physical medium a component is, generally the easier it is to determine such characteristics.

The performance of source recognition technologies depends on the type of recognition being attempted. Machine characteristics are generally more reliable than those imposed by human users of the systems.

Speaker Recognition

Speaker recognition is the process of estimating the identification of a speaker who is uttering voice signals [35]. The function can be divided into speaker identification and speaker verification as discussed above. Its use in the commercial sector is for automating such functions as bank transactions and voice mail. Interestingly enough humans have little difficulty in identifying speakers even when the signaling conditions are relatively poor.

Speaker identification can be classified as to whether it is text-dependent or text-independent. Text-dependent systems allow only a limited set of words to be used to perform the identification. Text-independent identification applies no such constraint. In general, text-dependent systems perform better than text-independent systems. Both requirements exist in communication EW systems. Text-independent capabilities apply when free and open speech is being processed, which applies in many circumstances. Text-dependent systems apply when the messages being processed are pro forma in character, such as artillery call for fire.

The language of the speaker may or may not be important. Humans can do a reasonably good job of recognizing other human voices, even over band-limited voice paths such as telephone channels. The objective behind automated speaker recognition technology is to have machines do the same thing.

In general, speaker identification applied to communication EW systems should be text-independent and robust. The latter means that performance must degrade smoothly as the SNR is decreased. There are cases that the former need not be true but those would be special situations.

Over relatively quiet channels, such as telephone lines, the problem is easier than over radio channels where noise and interference are prevalent. However, channel variability can be a problem over telephone lines as well. With the advent of digital telephony and fiber-optic telephone infrastructures, noise on telephone lines is rapidly becoming a problem of the past (in the developed countries as well as urban areas in undeveloped countries, anyway).

The speech spectrum is typically calculated with 20-ms time samples. Often the features used for speaker recognition are based on the cepstrum. The cepstrum contains information about time delays in speech. If frame represents a time segment of the voice signal, then

$$\text{cepstrum(frame)} = FFT^{-1}(\log|\text{FFT(frame)}|) \qquad (7.12C$$

Typically only the first several cepstral coefficients retained are required as features. The cepstrum is used to estimate the vocal tract parameters of the speaker. The subsequent computations after the cepstrum is obtained separate from the parameters of the speech that are related to the vocal tract from those that are related to the pitch information.

One model of the vocal process in humans has periodic pulses generated by forcing air through the vocal chords. The vocal tract filters these pulses and it is the parameters of the vocal tract that are useful in identifying one individual from another. These processes can be viewed as (short term) linear time-invariant processes, so if $P(f)$ represents the Fourier transform of the pulses generated by and emitted from the vocal chords, and if $T(f)$ represents the filter function of the vocal tract, then

$$S(f) = P(f)T(f) \qquad (7.121)$$

where $S(f)$ is the Fourier transform of the speech signal $s(t)$. In logarithms:

$$\log[S(f)] = \log[P(f)] + \log[T(f)] \qquad (7.122)$$

The periodic pulses produce peaks in the $\log[P(f)]$ spectrum while the characteristics of the vocal tract are in the second term, which is represented by variables that are changing more slowly and are therefore represented by the shape of the cepstral spectrum, as represented by the lower cepstral component values. In this process, the higher frequencies of the $\log|FFT|$ are suppressed because it is felt that the lower frequencies of human speech produce more useful information. The resultant spectrum is called the mel-cepstrum. The features generated this way make up a point in feature space. Ideally a single, well-defined point in this space represents each speaker, but ill-defined borders can yield confusing results.

Morse Code Operator Recognition

Morse code is the transmission of dots and dashes that represent coding of some language, English, for instance. In addition to those features associated with a transmitter used to send the code, features of the person operating the key can sometimes be used to identify the operator. A particular operator may have consistent durations for the dots and dashes, which are differentiable from other such operators. Particular sequences of symbols may have repeatable characteristics that can be used

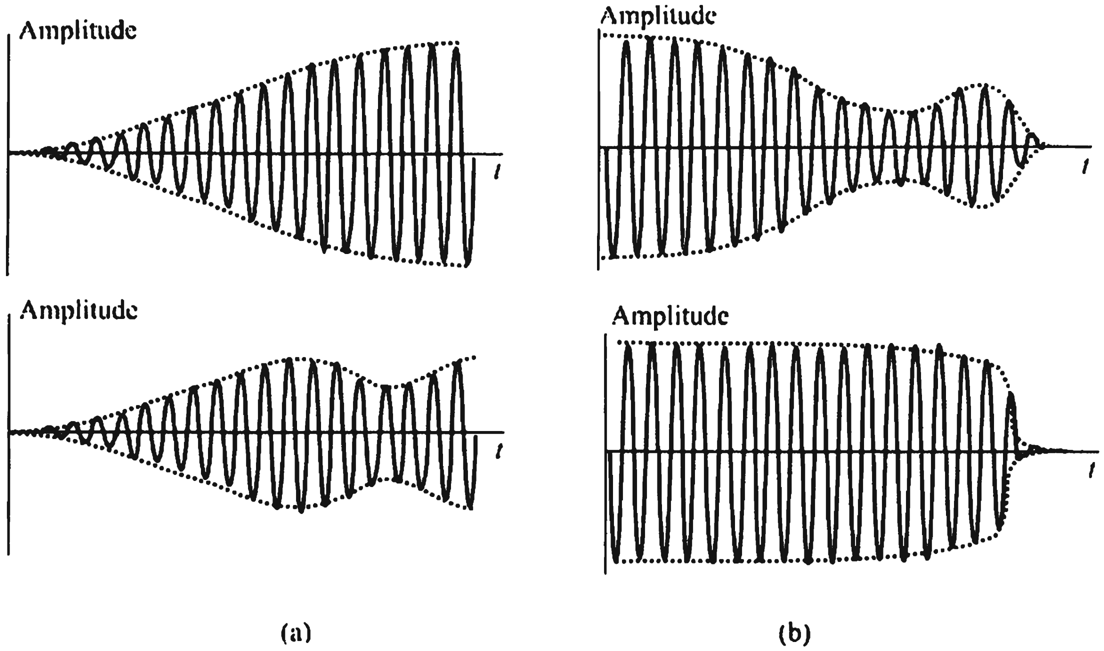


Figure 7.50 (a) Turn-on and (b) turn-off transients of a CW transmitter sending Morse code.

for identification. The character for *u* (• • -) would almost always follow *q* (• • - •) in the English language and this sequence may always be sent in a unique way (such as speed) by a particular Morse code key operator. Anther possible parameter is the duration of the code characters.

The key itself may have measurable features that can be used. In that case, of course, one is identifying the key and not necessarily the operator but this could be just as useful from an information point of view.

The transients in the amplitude of the RF carrier at turn-on and turn-off may be unique for a particular transmitter and sufficiently different among transmitters to be useful for identification. A few such cases are illustrated in Figure 7.50. The two turn-on transients are certainly different from one another as are the turn-off transients. If these characteristics are repeatable over time (the longer, the better), then they could be used to track targets. Frequently characteristics such as these are persistent until a repair is made to the system that changes whatever components are responsible for establishing this behavior.

Machine (Computer) Recognition

When computers are communicating via modems, the characteristics of the modems may present features that are usable for identification. The tone frequencies in FSK

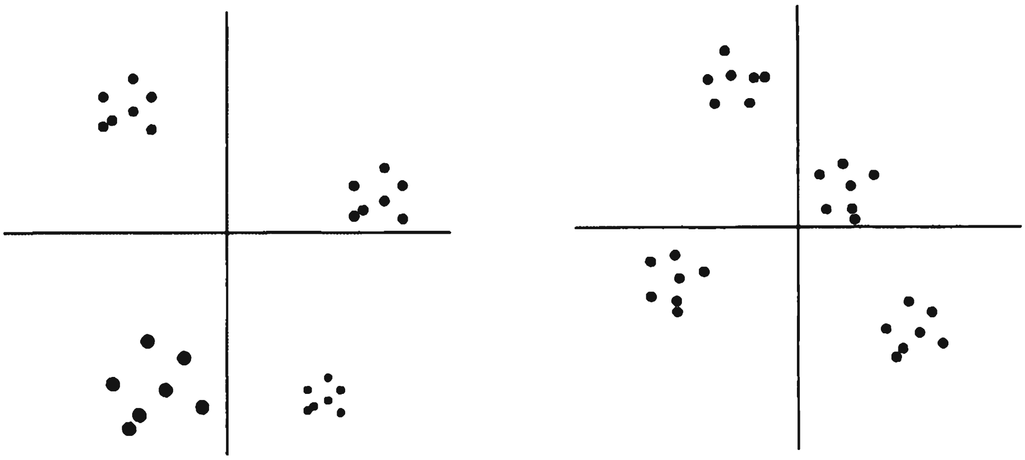


Figure 7.51 Two QAM or PSK constellations that exhibit different features that may be useful for modem identification.

modems are a simple example. These tones are not the same each time they are sent, but exhibit statistical properties that can be measured.

Phase shifts in PSK modems or phase shifts and amplitude variations in QAM modems also have statistical properties. Two such modems might have the constellations shown in Figure 7.51, from which features could be extracted for identification. For example, one such possibility might simply be the amplitude in this case. Another may be the distance between the centers of the constellation points.

Even when modems are not being used and the communication media is driven directly with digital signals (V.90, for example), the source may contain identifiable features. Timing jitter on system clocks may be unique to an individual or a class of machines. Baud rate properties may be usable Characteristic amplitude shifts may also contain features.

7.7.4.3 Language Identification

In those cases when the content of a transmission is important, and that transmission is a voice conversation, it may be necessary to first ascertain in which language the conversation is occurring [36]. Thus, automated language identification may be necessary. This is especially true in cases when automated translation is performed. Language identification as well as automated translation systems can be applied to written text as well. Herein this is not the case of interest.

In the commercial sector, language identification is important to, for example, phone companies so that calls can be routed to operators with the correct language. The same is true for international business transactions.

The features for language identification are usually based on the acoustic signature of that language. The acoustic signature is the unique characteristics of a language that make it different from all others. Possible sources of information to ascertain the acoustic signature are listed as follows:

- Acoustic phonetics: The tones characteristic of a language, which differ among languages, and, even when close, whose frequency of use varies;
- Prosodics: The duration of phones, speech rate, and intonation;
- Phonotactics: The rules that govern the use of phones in a language;
- Vocabulary.

Recent results indicate that 60–80% of correct language classification is possible with phone line quality sources of audio using these parameters. Phone line quality audio is usually very good, with a characteristically high SNR. Unfortunately, phone line quality speech is often unavailable, especially in wireless communications. RF transmitted voice is typically much noisier. Cellular phone and PCS phone audio is somewhere in between these two extremes. Because there is more data to work with, the longer the utterance, the better.

7.7.4.4 Emitter Identification

Discovering properties of a transmitter in order to ascertain what type of radio it is and, perhaps, what specific transmitter it is, is sometimes possible. Identification of the type of transmitter can assist in determining the EOB. Identifying a specific transmitter that has been intercepted in the past can assist in tracking battlespace units over time and space when that transmitter can be associated with a specific unit.

Type

Some of the parameters that can help identify the type of transmitter include the RF bandwidth, the frequency of operation, the type of modulation (e.g., AM, FM, and FSK), and the power level. With the advent of software programmable radios, which emulate the waveforms of other radios, identification of the type of radio will likely become more difficult over time.

The frequency of operation can be an indication of the purpose for which the transmitter was built. For example, reception of a signal in the low VHF frequency range is an indication that the radio is probably a military system. Reception of a signal in the 130 MHz range would indicate that the radio is probably intended for aircraft traffic control.

Identification of the type of modulation can be used to segregate transmitters by type. The modulation index typically varies somewhat even in the same radio type due to variations in the components used to make the modulator.

As for the transmitter sending Morse code above, the transients in PTT radio networks may also be usable as features. In the case of CW or AM transmitters, the features shown in Figure 7.50 may also apply to PTT networks. For the case of FM, the signal may have to be demodulated first but the demodulated output may exhibit features similar to those shown in Figure 7.50. Unintentional tones located anywhere in the spectrum may indicate particular oscillators in the transmitter, which are likely to be unique to a type of transmitter.

Specific

Once a transmitter has been detected and parameters extracted, it may be possible to identify that specific transmitter again if it is intercepted a second time. If an association can be made between that specific transmitter and the entity using it, then specific units or persons can be tracked over time. The automatic extraction of features would be highly desirable, as there would be too many emitters in an MRC for operators to do this. Such features would likely include fine measurements of the parameters mentioned above for emitter type identification, as well as others.

For MFSK, the statistical variation of the frequency tones can be determined. The natural way to do this is via the FFT discussed earlier. Some pulsed transmitters exhibit unintentional modulation on the pulses. This modulation could be low levels of AM or FM signals generated in the transmitter by a wide variety of mechanisms.

Filters in transmitters frequently exhibit ringing effects. Most wideband transmitters implement tuned filters at the output of the final amplifier stage and before the antenna. This is to conjugate match the amplifier to the antenna for the purpose of transferring maximum power to the antenna. These filters are typically reasonably narrowband and therefore have a reasonably high Q. High Q filters tend to ring, or oscillate. This ringing may be large enough to0 be detectable and useful as an identifying feature.

Background noise in voice PTT networks may contain artifacts of the surrounding area within which the radio is operating (inside a tank or inside a busy TOC as examples). While over extended time these characteristics are likely to change (background noise in a TOC, for example), for short temporal conditions they may be useful for tracking targets.

The spectral shape of the predetected signal is measurable and contains usable features. The selectivity of the radio is determined by how sharp the filters are in the modulation process. If the skirts on the spectrum of signals are sharp, it is an indication that the radio is selective as opposed to broad skirts, which might indicate another type of radio. In order to remove modulation effects, the spectrums used for this purpose would need to be averaged.

7.8 Concluding Remarks

This chapter introduces the reader to some fundamental concepts associated with signal processing in communication EW systems. Probably the most prolific signal processing task in these systems is calculating the Fourier transform, or its fast approximation. This transform is useful for determining the frequency content of signals and is fundamental to many of the signal detection functional requirements.

Noise is always present when processing communication signals. Some of the newer signal processing techniques such as the wavelet transform, cyclostationary processing, and high-order processing of signals can be used to reduce much of this noise, promising to make the signal detection function perform better.

Some of the many signal processing applications are discussed in this chapter as well. This list is certainly not all-inclusive, but it is presented to illustrate applications of some of the signal processing techniques.

References

[1] Stephens, J. P., "Advances in Signal Processing Technology for Electronic Warfare," *IEEE Aerospace and Electronic Systems Magazine*, November 1996, pp. 31–37.

[2] Bracewell, R. N., *The Hartley Transform*, Oxford, UK: Oxford University Press, 1986, p. 7.

[3] Rioul, O., and M. Vetterli, "Wavelets and Signal Processing," *IEEE Signal Processing Magazine*, October 1991, pp. 14–37.

[4] Bruce, A., D. Donoho, and H. Gao, "Wavelet Analysis," *IEEE Spectrum*, October 1996, pp. 26–35.

[5] Guillemain, P., and R. Kronland-Martinet, "Characterization of Acoustic Signals through Continuous Linear Time-Frequency Representations," *Proceedings of the IEEE*, Vol. 84, No. 4, April 1996, pp. 561–585.

[6] Ramchandran, K., M. Vetterli, and C. Herley, "Wavelets, Subband Coding, and Best Bases," *Proceedings of the IEEE*, Vol. 84, No. 4, April 1996, pp. 561–585.

[7] Wornell, G., "Emerging Applications of Multirate Signal Processing and Wavelets in Digital Communications," *Proceedings of the IEEE*, Vol. 84, No. 4, April 1996, pp. 561–585.

[8] Mallat, S., "Wavelets for Vision," *Proceedings of the IEEE*, Vol. 84, No. 4, April 1996, pp. 604–614.

[9] Schroder, P., "Wavelets in Computer Graphics," *Proceedings of the IEEE*, Vol. 84, No. 4, April 1996, pp. 561–585.

[10] Graps, A., "Introduction to Wavelets," accessed November 2007, http://www/arma.com/IEEEwave/IW_see_wave.html.

[11] Daubechies, I., "Orthonormal Bases of Compactly Supported Wavelets," *Communications on Pure and Applied Mathematics*, Vol. 41, 1998, pp. 909–996.

[12] Bock, R. K., "Fast Transforms," *The Data Analysis Briefbook*, accessed March 2008, http://ikpe1101.ikp.kfa-juelich.de/briefbook_data_analysis/node83.html.

[13] Brigham, E. O., *The Fast Fourier Transform*, Englewood Cliffs, NJ: Prentice-Hall, 1974, Appendix A.

[14] Vaughn, R. G., N. L. Scott, and D. R. White, "The Theory of Bandpass Sampling," *IEI Transactions on Signal Processing*, Vol. 39, No. 9, September 1991, pp. 1973–1984.

[15] Gardner, W. A., "Signal Interception: A Unifying Theoretical Framework for Featur Detection," *IEEE Transactions on Communications*, Vol. 36, No. 8, August 1988, pp. 897–906.

[16] Mendel, J. M., "Tutorial on Higher-Order Statistics (Spectra) in Signal Processing and System Theory: Theoretical Results and Some Applications," *Proceedings of the IEEE*, Vol. 79, No. 3, March 1991, pp. 278–305.

[17] Kostylev, V. I., "Energy Detection of a Signal with Random Amplitude," *Proceedings IEEE International Conference on Communications*, 2002, April 28–May 2, Vol. 3, pp. 1606–1610.

[18] Root, W. L., "An Introduction to the Theory of the Detection of Signals in Noise," *Proceedings of the IEEE*, Vol. 58, No. 5, May 1970, pp. 610–623.

[19] Sonninschein, A., and P. M. Fishman, "Radiometric Detection of Spread-Spectrum Signals in Noise of Uncertain Power," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 28, No. 3, July 1992, pp. 654–660.

[20] Mills, R. E., and G. E. Prescott, "A Comparison of Various Radiometer Detection Models," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 32, No. 1, January 1996, pp. 467–473.

[21] Torrieri, D. J., *Principles of Secure Communication Systems*, 2nd Ed., Norwood, MA: Artech House, 1992, pp. 294–307.

[22] Woodring, D. G., "Performance of Optimum and Suboptimum Detectors for Spread Spectrum Waveforms," NRL Report 8432, December 30, 1980.

[23] Levitt, B. K., and U. Cheng, "Optimum Detection of Frequency-Hopped Signals," *Proceedings IEEE MILCOM*, 1992.

[24] Levitt, B. K., et al., "Optimum Detection of Slow Frequency-Hopping Signals," *IEEE Transactions on Communications*, Vol. 42, Nos. 2/3/4, February/March/April 1994, p. 1990.

[25] Beaulieu, N. C., W. L. Hopkins, and P. J. McLane, "Interception of Frequency-Hopped Spread-Spectrum Signals," *IEEE Journal on Selected Areas of Communications*, Vol. 8, No. 5, June 1990, pp. 854–855.

[26] Rifkin, R., "Comparison of Performance Measures for Intercept Detectors," *Proceedings IEEE Tactical Communications Conference, 1994, Digital Technical for the Tactical Communicator*, pp. 510–517.

[27] Nandi, A. K., and E. E. Azzouz, "Algorithms for Automatic Modulation Recognition of Communication Signals," *IEEE Transactions on Communications*, Vol. 46, No. 4, April 1998, pp. 431–436.

[28] Azzouz, E. E., and A. K. Nandi, "Procedure for Automatic Recognition of Analogue and Digital Modulations," *IEE Proceedings—Communications*, Vol. 143, No. 5, October 1996, pp. 259–266.

[29] Azzouz, E. E., and A. K. Nandi, *Automatic Modulation Recognition of Communication Signals*, Boston, MA: Kluwer Academic Publishers, 1996.

[30] Assaleh, K., K. Farrell, and R. J. Mammone, "A New Method of Modulation Classification for Digitally Modulated Signals," *Proceedings IEEE MILCOM*, 1992, pp. 712–716.

[31] Whelchel, J. E., et al., "Signal Understanding: An Artificial Intelligence Approach to Modulation Classification," *Proceedings IEEE MILCOM*, 1984, pp. 231–236.

[32] Hush, D. R., and B. G. Horne, "Progress in Supervised Neural Networks," *IEEE Signal Processing Magazine*, January 1993.

[33] Kim, K., and A. Polydoros, "Digital Modulation Classification: The BPSK versus QPSK Case," *Proceedings IEEE MILCOM*, 1988, pp. 431–436.

[34] Sills, J. A., "Maximum-Likelihood Modulation Classification for PSK/QAM," *Proceedings IEEE MILCOM*, 1999.

[35] Gish, H., and M. Schmidt, "Text-Independent Speaker Identification," *IEEE Signal Processing Magazine*, October 1994, pp. 18–32.

[36] Muthusamy, Y. K., E. Barnard, and R. A. Cole, "Reviewing Automatic Language Identification," *IEEE Signal Processing Magazine*, October 1994, pp. 33–41.

Chapter 8

Direction-Finding Position-Fixing Techniques

8.1 Introduction

Determining the geolocation of an emitting target, as mentioned previously, is an important aspect of ES systems. This is more commonly known as *position fixing*. There are several methods to accomplish this and some of them are described in this chapter. In all cases some parameters associated with the incoming signal, such as its time of arrival or differential time of arrival measured at two or more locations, are used to compute the location of the emitter.

8.2 Bearing Estimation

One technique for position fixing determines the *angle of arrival* (AOA) or LOB of the incoming signal relative to magnetic or true north (they are not the same) at two or more locations. Where these bearings intersect is taken as the location of the target. This is sometimes referred to as triangulation. Herein, these bearings will be referred to as LOPs.

To determine the LOPs, several techniques can be used. The phase (or almost equivalently, time) difference can be measured between the signals at two or more antenna elements. Alternately if the antennas have directionality, the amplitude difference between these two signals can be measured. Frequency differences measured between two antennas can be used to determine the bearing, if one or more of the antennas are moving relative to the other or to the target. Some of these techniques are discussed in this section.

LOP systems normally do not operate at the frequencies of the signals. The frequency is typically converted to an IF and the phase or time measurements are made on this converted signal. Ignoring noise and other error sources, the IF

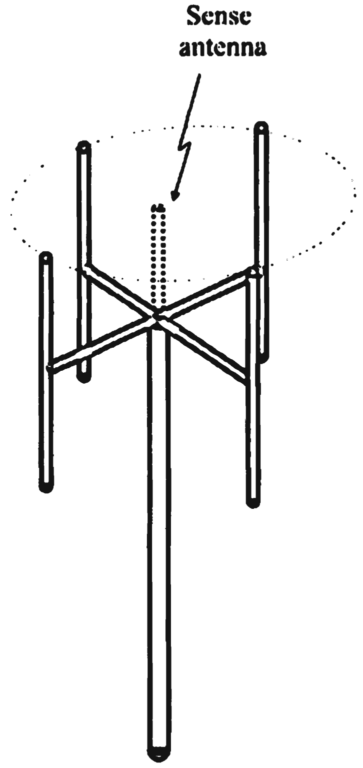


Figure 8.1 A square antenna array also forms a circular antenna array. The circle shown here is not part of the antenna but is included to show the circular nature of the square array.

version of the RF signals has the same information in it as the latter—it is just (usually) lower in frequency. Hereafter, the signals at RF and IF will be used interchangeably. If the difference is important, it will be so noted.

8.2.1 Circular Antenna Array

One of the more common forms of antenna arrays for bearing determination is a circular array, commonly known as a *circularly disposed antenna array* (CDAA). An example of a four-element array mounted on a mast is illustrated in Figure 8.1. Other forms of this antenna include more or fewer numbers of elements. A sense antenna is sometimes included with a circular array, which is used to remove ambiguities in the bearing. This sense antenna can be formed virtually by combining in phase the output of the other antenna elements [1].

Let R be the radius of the CDAA. Thus, R is the length of an "arm" of the array measured from the center to any one of the antenna elements. The quantity $2R/\lambda$ is sometimes referred to as the *aperture* of a circular array and it indicates the number of wavelengths presented by the array to a signal impinging orthogonal to the plane of the array which is the plane determined by the arms connecting the antenna elements.

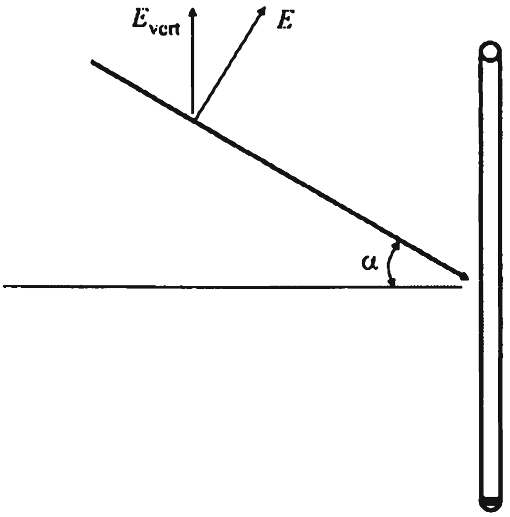


Figure 8.2 An E-field of arbitrary orientation incident on a vertically orientated antenna will only excite the antenna with the vertical component of the E-field.

An incoming signal $s(t)$ represented in phasor form is shown in Figure 8.2. The EM wave, as indicated previously, has an associated E-field $E(t)$ with a magnitude E. The vertical component of this field is the only portion that excites a vertically oriented antenna. Thus, E_{vert} is given by E cos.α and the corresponding signal amplitude must be adjusted by this factor.

8.2.2 Interferometry

One popular technique for measuring the AOA of an incoming signal, and thereby determining its LOP, is *interferometry*. In this case the phase difference $\Delta\phi_{12}$ or time difference Δt_{12} between two antennas is measured directly by some means. These two types of measurement produce the same result, the major difference being what is measured.

It is also possible to implement an active interferometer [2]. Active interferometers emit a signal and measure the phase or time difference of the returned signal.

There are two fundamental types of interferometers. The first measures the phase differences between two antennas and the angle of arrival is deduced from this angle. This type of interferometer is called a phase interferometer. The second type measures the time of arrival difference of the signal at two antennas. To make this measurement there must be some feature associated with the signal that can be used as a time mark, such as the leading edge of a radar pulse. Lacking such a feature the signals at the two antennas must be correlated against each other. Where this correlation peaks is an indication of the time difference. This is called

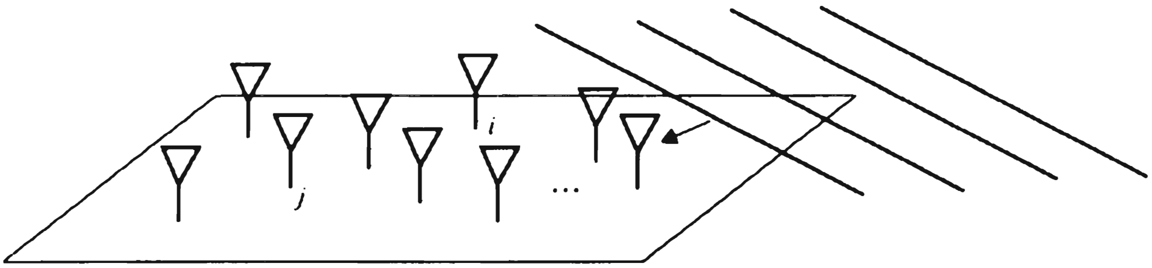


Figure 8.3 An arbitrary array of *M* sensors.

arrival time interferometry. We will now discuss the concepts involved with phase interferometry.

8.2.2.1 Phase Interferometry

Consider the arbitrary planar antenna array of *M* sensors illustrated in Figure 8.3. If *s*(*t*) represents the signal emitted from a far-field source then the outputs from *M* narrowband receiver channels connected to the output of these antennas at each time $t \in \{1, 2, \ldots, K\}$, can be represented as

$$\mathbf{r}(t) = \mathbf{a}(\theta)s(t) + \mathbf{n}(t), \qquad t = 1, 2, \ldots, K \tag{8.1}$$

where **r**(*t*) is an *M* × 1 vector with complex elements, *K* is the number of samples taken (snapshots), and **n**(t) is an *M* × 1 vector of uncorrelated antenna AWGN samples. **a**(θ) is the *M* × 1 complex vector representing the phase response of the array with a signal impinging on it from direction θ.

An interferometer measures the phase differences from pair-wise antenna elements. Usually one is designated as the reference antenna and the phase differences are computed from each of the other elements compared to that reference. In general, the estimated phase difference, $\hat{\phi}_{ij}$, between the *i*th element and *j*th element is found from

$$\hat{\phi}_{ij} = \arg\left(\frac{1}{K}\sum_{t=1}^{K} r_i(t) r_j^*(t)\right) \tag{8.2}$$

The signal impinging on a second antenna displaced from an antenna as shown in Figure 8.4 must travel an extra distance of $d = D\sin\varphi$ compared with that antenna. This imposes an additional phase shift on the signal given by

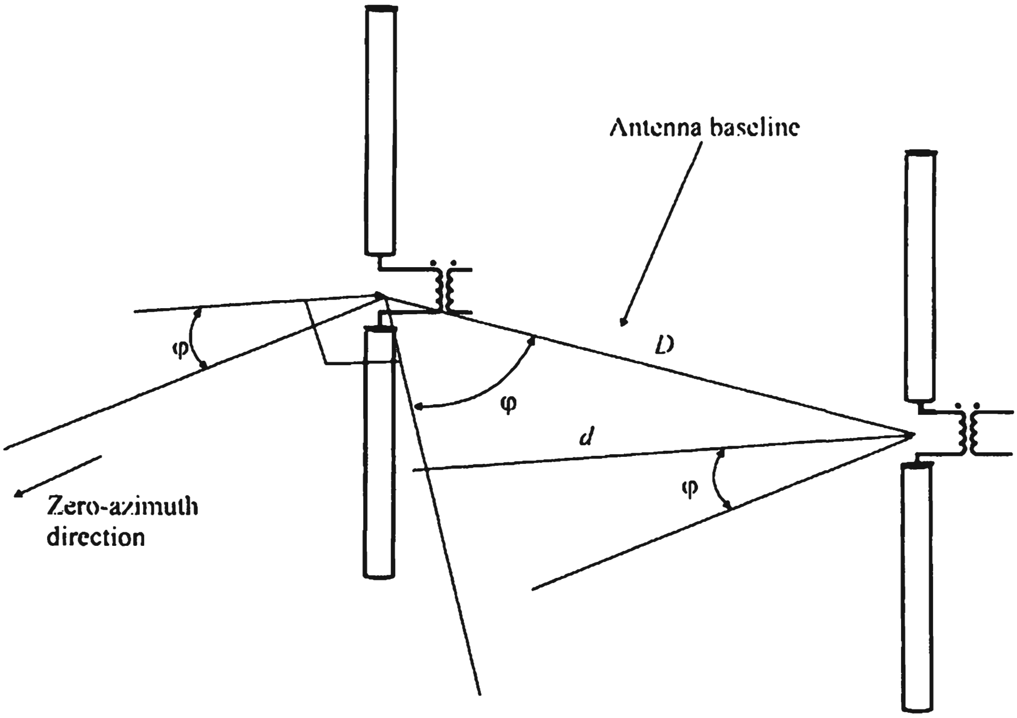


Figure 8.4 Definition of variables for an interferometer.

$$\phi = \frac{2\pi D}{\lambda}\sin\varphi \tag{8.3}$$

To determine how errors in calculation of the phase angle relate to errors in calculating the azimuth of arrival of the signal, consider the following [3]. If U represents some system parameter, which is a function of independent measurements represented by the variables, u_1, u_2, ..., u_N and the u_i have probable errors denoted by $e(u_i)$, then the probable error of U is approximated by

$$e(U) = \sqrt{\left(\frac{\partial U}{\partial u_1}\right)^2 e^2(u_1) + \left(\frac{\partial U}{\partial u_2}\right)^2 e^2(u_2) + \cdots + \left(\frac{\partial U}{\partial u_N}\right)^2 e^2(u_N)} \tag{8.4}$$

If the probable errors are from an independent statistical set with zero means, then the probable error is the same as the standard deviation $e(u_i) = \sigma_i$. In the simplified case under consideration here, there is one independent variable, φ. Thus in this expression, $U = \phi$ and $u_1 = \varphi$. Assume that the mean of ϕ is zero so that

$$\sigma_\phi = \sqrt{\left(\frac{\partial\phi}{\partial\varphi}\right)^2 \sigma_\varphi^2} = \frac{\partial\phi}{\partial\varphi}\sigma_\varphi \tag{8.5}$$

The partial derivative in this expression can be calculated from (8.3) as

$$\frac{\partial\phi}{\partial\varphi} = \frac{2\pi D}{\lambda}\cos\varphi \tag{8.6}$$

which yields the expression for the standard deviation of the azimuth calculation as

$$\sigma_\phi = \frac{2\pi D}{\lambda}\cos\varphi\sigma_\varphi \tag{8.7}$$

Thus, the standard deviation of the azimuth measurement is maximum when $\varphi = 0$, or off the ends of the baseline. It is minimum for $\varphi = \pi/2$, or orthogonal to the baseline.

There are theoretical limits on how accurately parameters associated with signals can be measured when noise is present and taken into consideration. Herein the one that will be used is called the *Cramer-Rao* lower bound and it is a measure of parameter estimation accuracy in the presence of white noise. White noise is an approximation in most situations and can be a fairly good approximation. For example, the hissing noise that one hears sometimes from a radio when it is not tuned to a station, especially when the volume is turned up, is very close to white noise. The term "white" comes from the fact that white light contains all of the colors—white noise contains energy from all of the frequencies.

For an interferometric line of position system, the Cramer-Rao bound σ can be calculated to be [4]

$$\sigma^2 = \left(\frac{E}{N_0}\right)^{-1} \tag{8.8}$$

where E is the signal energies at the two antennas (watt-seconds), assumed to be the same, and N_0 is the noise spectral density at the two antennas (watts per hertz), also assumed to be the same and independent of each other. The other assumption for this equation is that the angle of arrival, θ, does not vary over the measurement interval.

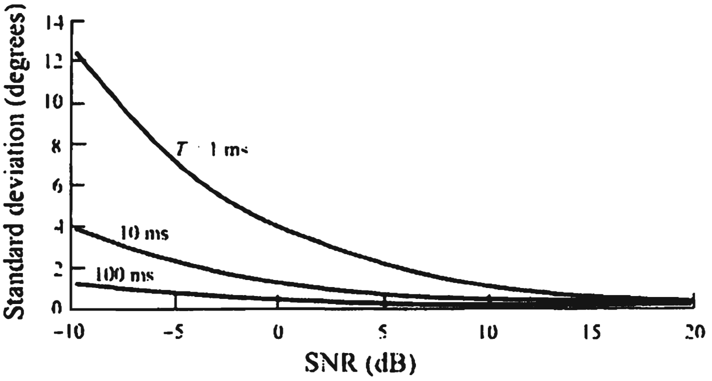


Figure 8.5 Cramer-Rao interferometric bound for $B = 200$ kHz.

To convert this equation into a more usable form, recall that $E = S$ (W) T (seconds), and that N (W) = N_0 (W/Hz) B (Hz), where N is the noise power and B is the noise bandwidth, the latter assumed for convenience here to be the IF bandwidth. After some trivial algebra,

$$\sigma^2(\varphi) = \left(\frac{S}{N}TB\right)^{-1} \tag{8.9}$$

with $\sigma(\varphi)$ in radians. Therefore, the variance is reduced (accuracy improved) by increasing the SNR, the integration time, or the measurement bandwidth. Note, however, that increasing the bandwidth beyond the bandwidth of the signal of interest will decrease the SNR, so care must be used. With these assumptions, and an assumed noise bandwidth of 200 kHz, the best instrumental accuracy possible, as given by the standard deviation of the measurement of the LOP, is shown versus (power) SNR and various measurement times in Figure 8.5.

8.2.2.2 Correlation

In this method, the correlation is computed between the phase difference samples between two antennas compared with stored (calibrated) phase difference responses between these two antennas [5–8]. The correlation function is given by

$$r_{xy}(i) = \frac{\phi_{\text{m}}^{\text{T}}\phi_{\text{R}}(i)}{\sqrt{\phi_{\text{m}}^{\text{T}}\phi_{\text{m}}\phi_{\text{R}}^{\text{T}}(i)\phi_{\text{R}}(i)}} \tag{8.10}$$

where $\vec{\phi}_m$ is the measured phase difference vector and $\vec{\phi}_R(i)$ is the ith calibrated phase difference vector. The angle where (8.10) is maximized is used as the estimate of the correct phase difference.

8.2.2.3 Least-Squared Error

In the least-squared-error method, the response vector (8.1) is compared with the stored response vectors and selects the one that minimizes the vector difference. The LSE objective function is given by

$$g(\theta) = \mathbf{e}^{\mathrm{T}}(\theta)\mathbf{R}_{\mathrm{LSE}}^{-1}\mathbf{e}(\theta) \tag{8.11}$$

where, for $M = 4$ for example,

$$e(\theta) = \left[\hat{\phi}_{13} - \phi_{13}(\theta), \cdots, \hat{\phi}_{41} - \phi_{41}(\theta)\right]^{\mathrm{T}} \tag{8.12}$$

$\hat{\theta}$ is chosen such that

$$g(\hat{\theta}) = \max_{\theta} g(\theta) \tag{8.13}$$

8.2.2.4 Triple-Channel Interferometer

The antenna configuration shown in Figure 8.6 may be used as a three-channel interferometer. There are three baselines established by the three antenna elements, one for each antenna pair. Figure 8.7 shows the top view of this antenna arrangement. The AOA is related to the antenna parameters as follows.

The distance traveled by a signal impinging on this array from an azimuth angle denoted by θ and an elevation angle denoted by α, between antenna 0 and the center of the array, given by d_{0c} is calculated as

$$d_{0c} = R\cos\alpha\cos\varphi \tag{8.14}$$

The cos α term is required to reflect the signal perpendicular to the plane of the antenna array as mentioned above. The phase difference corresponding to this distance is given by

$$\phi_{0c} = 2\pi f \Delta t_{0c}$$

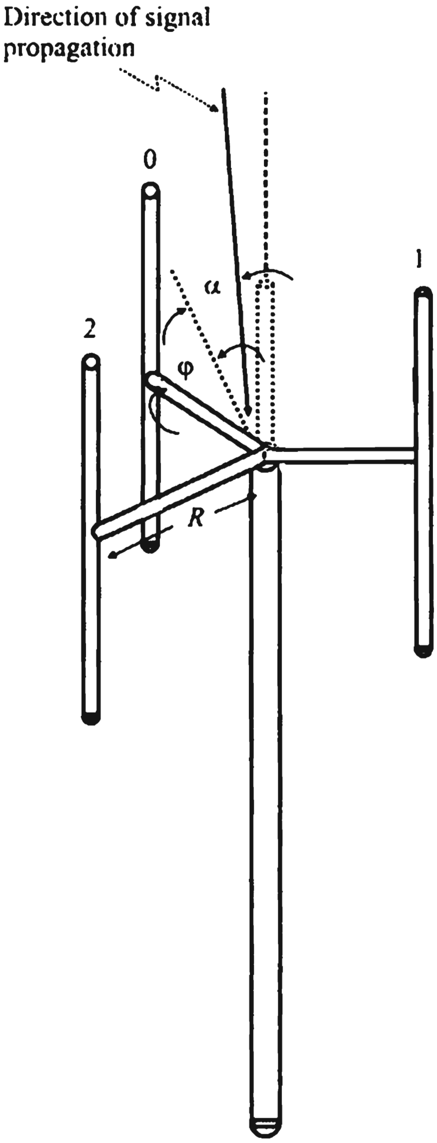


Figure 8.6 Definition of variables for a triple-channel interferometer.

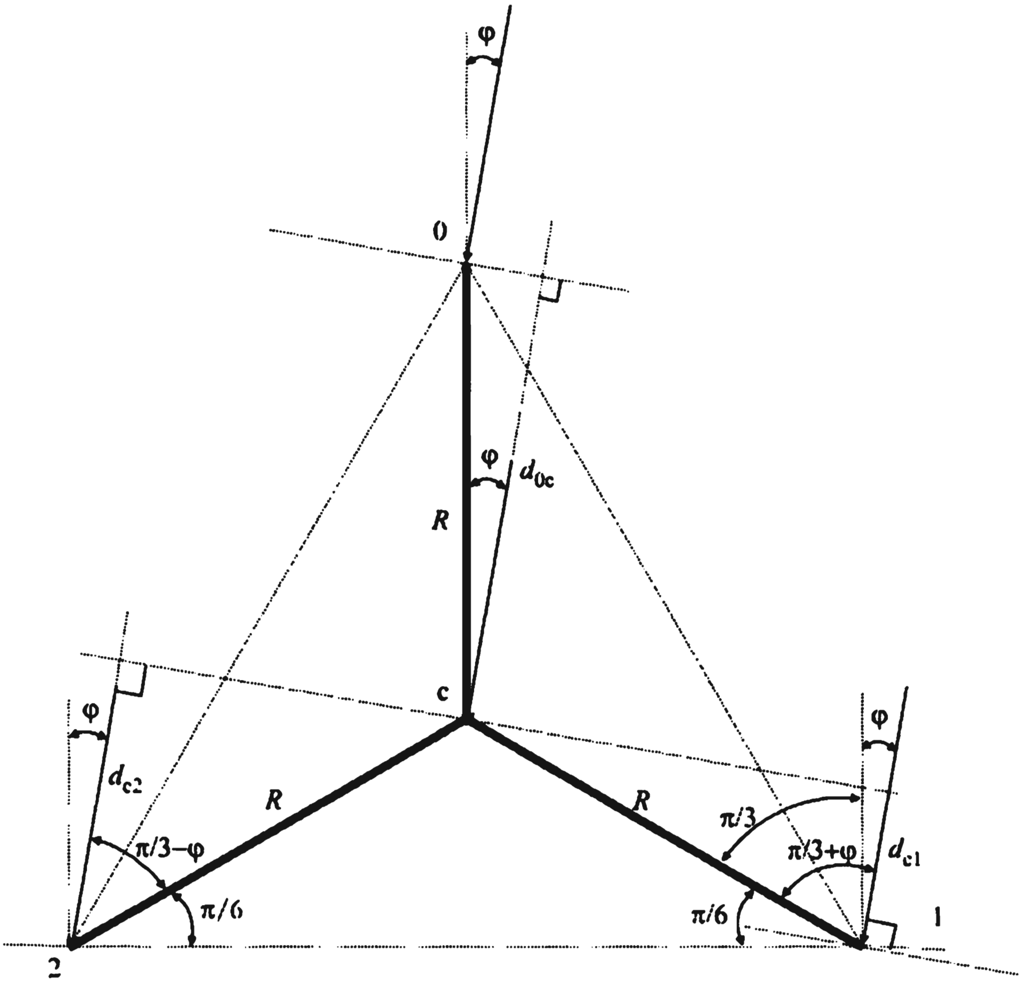


Figure 8.7 Top view of a triple-channel interferometer. Shown are the definitions for the angles and wave fronts as they pass through the array.

$$= 2\pi \frac{c}{\lambda} \Delta t_{0c} \tag{8.15}$$

where Δt_{0c} is the time it takes to traverse this distance. But $d_{0c} = c\Delta t_{0c}$ so

$$\begin{aligned} \phi_{0c} &= 2\pi \frac{d_{0c}}{\lambda} \\ &= \frac{2\pi R}{\lambda} \cos\alpha \cos\varphi \end{aligned} \tag{8.16}$$

Likewise, the distance traveled by this same signal from the center of the array to antenna 1 is given by

$$d_{c1} = R\cos\alpha\cos\left(\frac{\pi}{3}+\varphi\right) \tag{8.17}$$

with a corresponding phase difference given by

$$\phi_{c1} = \frac{2\pi R}{\lambda}\cos\alpha\cos\left(\frac{\pi}{3}+\varphi\right) \tag{8.18}$$

The minus sign is required here because the phase at antenna 1 lags behind the phase at the center of the array.

The distance traveled between the center of the array and antenna 2 is given by

$$d_{c2} = R\cos\alpha\cos\left(\frac{\pi}{3}-\varphi\right) \tag{8.19}$$

with the phase difference being

$$\phi_{c2} = \frac{2\pi R}{\lambda}\cos\alpha\cos\left(\frac{\pi}{3}-\varphi\right) \tag{8.20}$$

The signals at each of the antennas can thus be represented in terms of the AOA and elevation angle as follows.

$$s_0(t) = s(t)\cos\alpha e^{j\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi} \tag{8.21}$$

$$s_1(t) = s(t)\cos\alpha e^{-j\frac{2\pi R}{\lambda}\cos\alpha\cos\left(\frac{\pi}{3}+\varphi\right)} \tag{8.22}$$

$$s_2(t) = s(t)\cos\alpha e^{-j\frac{2\pi R}{\lambda}\cos\alpha\cos\left(\frac{\pi}{3}-\varphi\right)} \tag{8.23}$$

To calculate the AOA, φ, and the elevation angle, α, first define $\Delta_1 = \phi_{c2} - \phi_{c1}$. Then

$$\Delta_1 = -\frac{2\pi}{\lambda}R\cos\alpha\left[\cos\left(\frac{\pi}{3}-\varphi\right)+\cos\left(\frac{\pi}{3}+\varphi\right)\right]$$

$$
\begin{aligned}
&= -\frac{2\pi}{\lambda} R\cos\alpha \left[\begin{matrix} \cos\left(\frac{\pi}{3}\right)\cos\varphi + \sin\left(\frac{\pi}{3}\right)\sin\varphi \\ \qquad + \cos\left(\frac{\pi}{3}\right)\cos\varphi - \sin\left(\frac{\pi}{3}\right)\sin\varphi \end{matrix} \right] \\
&= -\frac{2\pi}{\lambda} R\cos\alpha \left[2\cos\left(\frac{\pi}{3}\right)\cos\varphi \right] \\
&= \frac{2\pi}{\lambda} R\cos\alpha\cos\varphi
\end{aligned} \tag{8.24}
$$

Define $\Delta_2 = \phi_{0c} + \phi_{c1}$. Then

$$
\begin{aligned}
\Delta_2 &= \frac{2\pi}{\lambda} R\cos\alpha \left[\cos\theta - \cos\left(\frac{\pi}{3} + \varphi\right) \right] \\
&= \frac{2\pi}{\lambda} R\cos\alpha \left[\cos\theta - \cos\left(\frac{\pi}{3}\right)\cos\varphi + \sin\left(\frac{\pi}{3}\right)\sin\varphi \right] \\
&= \frac{2\pi}{\lambda} R\cos\alpha \left[\frac{1}{2}\cos\varphi + \frac{\sqrt{3}}{2}\sin\varphi \right]
\end{aligned} \tag{8.25}
$$

The AOA is given by

$$
\varphi = \tan^{-1}\left(\frac{\phi_{0c}}{\Delta_1} \right) \tag{8.26}
$$

because

$$
\begin{aligned}
\tan^{-1}\left(-\frac{\phi_{0c}}{\Delta_1} \right) &= \tan^{-1}\left(-\frac{\frac{2\pi}{\lambda} R\cos\alpha\sin\varphi}{-\frac{2\pi}{\lambda} R\cos\alpha\cos\varphi} \right) \\
&= \tan^{-1}\left(\frac{\sin\varphi}{\cos\varphi} \right)
\end{aligned} \tag{8.27}
$$

To obtain an expression for the elevation angle, expand

$$
\begin{aligned}
&\left(\frac{\sqrt{3}}{2}\Delta_1\right)^2+\left(\Delta_2-\frac{1}{2}\Delta_1\right)^2\\
&\quad=\frac{3}{4}\left(\frac{2\pi R}{\lambda}\right)^2\cos^2\alpha\cos^2\varphi+\left(\frac{2\pi R}{\lambda}\cos\alpha\right)^2\\
&\qquad\times\left(\frac{1}{2}\cos\varphi+\frac{\sqrt{3}}{2}\sin\varphi-\frac{1}{2}\cos\varphi\right)^2\\
&\quad=\frac{3}{4}\left(\frac{2\pi R}{\lambda}\right)^2\cos^2\alpha\cos^2\varphi+\left(\frac{2\pi R}{\lambda}\cos\alpha\right)^2\left(\frac{3}{4}\sin^2\varphi\right)\\
&\quad=\frac{3}{4}\left(\frac{2\pi R}{\lambda}\right)^2\cos^2\alpha(\cos^2\varphi+\sin^2\varphi)\\
&\quad=\frac{3}{4}\left(\frac{2\pi R}{\lambda}\right)^2\cos^2\alpha
\end{aligned}
$$

so that

$$
\alpha=\cos^{-1}\sqrt{\frac{4}{3}\left(\frac{\lambda}{2\pi R}\right)^2\left[\left(\frac{\sqrt{3}}{2}\Delta_1\right)^2+\left(\Delta_2-\frac{1}{2}\Delta_1\right)^2\right]} \tag{8.28}
$$

Thus, by measuring the phase differences between the antenna elements, the AOA in the plane of the antenna as well as the elevation AOA of a signal can be determined with a triple channel interferometer. These measurements are dependent on the array parameter, R, as well as the frequency (via the wavelength) of the signal. The larger the array radius, R, relative to the wavelength of the signal, the better these AOAs can be measured. There is a limit on this length, however. The baseline length, which in this case is along a diagonal of the periphery of the array, must be less than one-half of the wavelength to avoid ambiguous results. If it is longer than this, then multiple AOAs will generate the same time differences and therefore the same indicated angle. The wavelength is smallest at the highest operating frequency of the antenna array, so at the lowest operating frequency we expect the poorest measurement accuracy.

8.2.2.5 Four-Element Interferometer

The geometric relationships in this antenna system are as shown in Figure 8.8 with a top view shown in Figure 8.9. As above, R is the radius of the array while the

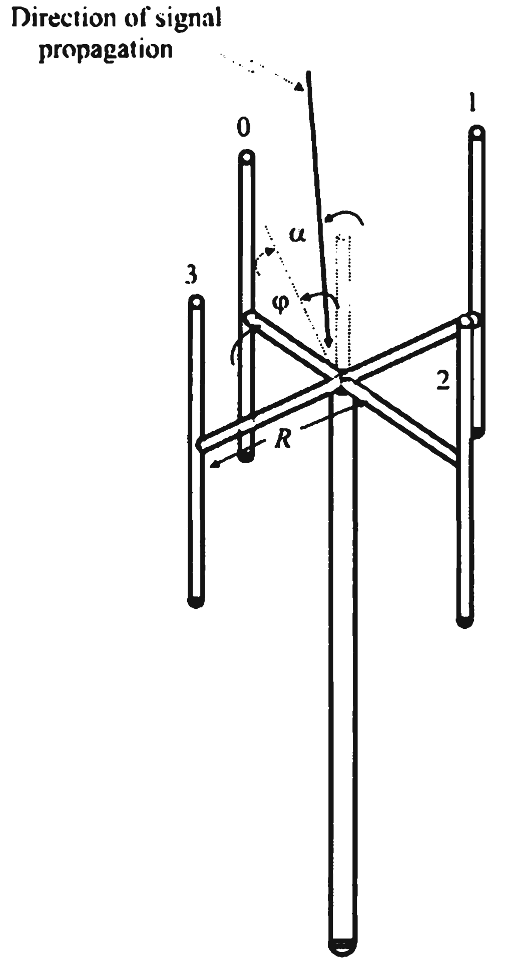


Figure 8.8 Defining the terms associated with a four-element circular array.

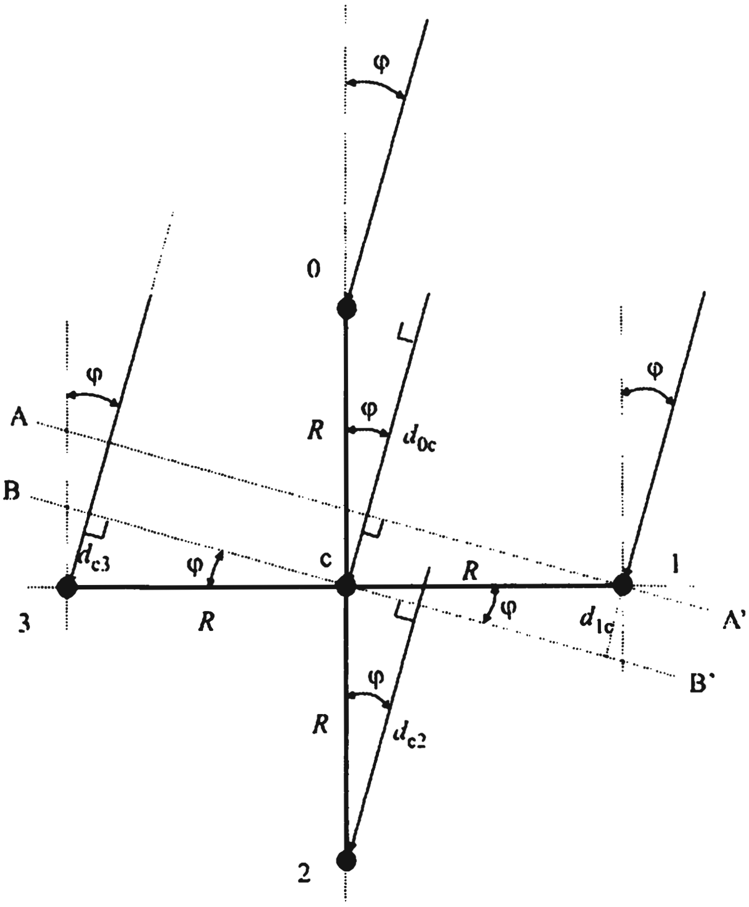


Figure 8.9 Angles and phase lines looking down from the top of a four-element circular array.

azimuth AOA of the signal $s(t)$ is given as θ while the arrival angle of the signal in the vertical dimension (zenith) is α relative to the plane of the array. The frequency of the signal is given as $f = c/\lambda$, the number of antenna elements $N = 4$, and c is the speed of light. It is assumed that $R < \lambda/4$ so that phase ambiguities do not arise.

If $s(t)$ represents the signal at the center of the array, then at each antenna element a replica of $s(t)$ exists but has shifted in phase somewhat. Thus

$$s_0(t) = s(t)\cos\alpha e^{j\cos(\alpha)\phi_0} \tag{8.29}$$

$$s_1(t) = s(t)\cos\alpha e^{j\cos(\alpha)\phi_1} \tag{8.30}$$

$$s_2(t) = s(t)\cos\alpha e^{j\cos(\alpha)\phi_2} \tag{8.31}$$

$$s_3(t) = s(t)\cos\alpha e^{j\cos(\alpha)\phi_3} \tag{8.32}$$

where ϕ_i's represent this phase shift. As above, the cos α multiplying terms on $s(t)$ project the magnitude of the incoming signal $s(t)$ perpendicular to the array plane since the elevation angle of the signal relative to this plane is α. In Figure 8.9 the phase difference at antenna element 1 can be calculated as follows.

Now

$$\cos\alpha\sin\varphi = \frac{d_{1c}}{R} \tag{8.33}$$

where d_{1c} is the distance traveled by the signal iso-phase line A-A′ to iso-phase line B-B′ in the array plane. Since

$$\begin{aligned}\phi_{1c} &= 2\pi f\Delta t_{1c}\\ &= \frac{2\pi c\Delta t_{1c}}{\lambda}\\ &= \frac{2\pi\cos\alpha d_{1c}}{\lambda}\\ &= \frac{2\pi R}{\lambda}\cos\alpha\sin\varphi\end{aligned} \tag{8.34}$$

but

$$\sin\beta = \cos(\beta - \frac{\pi}{2}) \tag{8.35}$$

for any β so that

$$\phi_{1c} = \frac{2\pi R}{\lambda}\cos\alpha\cos\left(\varphi - \frac{\pi}{2}\right) \tag{8.36}$$

The distance the signal must travel between the center of the array and antenna 2 is given by

$$d_{c2} = R\cos\alpha\cos\varphi \tag{8.37}$$

The corresponding phase shift associated with this distance is

$$\phi_{c2} = -\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi \tag{8.38}$$

The minus sign is required because the phase at antenna 2 lags that at the center of the array, whereas at antenna 1 above, the phase at antenna 2 led the phase at the center. These signs could be reversed as long as consistency is maintained.

Likewise, the distance the signal must travel between the center of the array and antenna 3 is given by

$$d_{c3} = R\cos\alpha\sin\varphi \tag{8.39}$$

while the corresponding phase shift is given by

$$\begin{aligned}\phi_{c3} &= -\frac{2\pi R}{\lambda}\cos\alpha\sin\varphi \\ &= -\frac{2\pi R}{\lambda}\cos\alpha\cos\left(\varphi - \frac{\pi}{2}\right)\end{aligned} \tag{8.40}$$

where, again, the phase at antenna 3 lags that at the center of the array, so that the minus sign is required.

The distance the signal must travel between antenna 0 and the center of the array is given by

$$d_{0c} = R\cos\alpha\cos\varphi \tag{8.41}$$

so that

$$\phi_{0c} = \frac{2\pi R}{\lambda}\cos\alpha\cos\varphi \tag{8.42}$$

Equations (8.29)–(8.42) can be summarized with (8.43) and (8.44) as

$$s_i(t) = s(t)\cos\alpha e^{j\frac{2\pi R}{\lambda}\cos\alpha\cos(-\frac{2\pi i}{N}+\varphi)} \quad i = 1,2,3,4 \tag{8.43}$$

and

$$\phi_i = \frac{2\pi R}{\lambda}\cos(-\frac{2\pi i}{N}+\varphi) \quad i = 1,2,3,4 \tag{8.44}$$

To determine the bearing calculate

$$\frac{\phi_{1c}}{\phi_{0c}} = \frac{\frac{2\pi R}{\lambda}\cos\alpha\sin\varphi}{\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi} \tag{8.45}$$

so that

$$\varphi = \tan^{-1}\left(\frac{\phi_{1c}}{\phi_{0c}}\right) \tag{8.46}$$

This calculation yields $0 \le \varphi < \pi$ so ϕ_{c2} and ϕ_{c3} are used to determine from which side of the array the signal is coming.

To determine the elevation formulate

$$\begin{aligned}\phi_{0c}^2 + \phi_{1c}^2 &= \left(\frac{2\pi R}{\lambda}\right)^2 \cos^2\alpha\cos^2\varphi + \left(\frac{2\pi R}{\lambda}\right)^2 \cos^2\alpha\cos^2\varphi \\ &= \left(\frac{2\pi R}{\lambda}\right)^2 \cos^2\alpha\left(\cos^2\varphi + \sin^2\varphi\right)\end{aligned} \tag{8.47}$$

yielding

$$\alpha = \sqrt{\cos^{-1}\left[\left(\frac{\lambda}{2\pi R}\right)^2 \left(\phi_{0c}^2 + \phi_{1c}^2\right)\right]} \tag{8.48}$$

So, again, like the triple-channel interferometer, a four-element circular antenna array can be used to measure the AOA of a signal, and thus facilitate calculation of the line of position to the target. It also allows calculation of the elevation AOA.

8.2.3 Monopulse Direction Finder

By shifting the phase of a signal that has been received by several antennas in specified ways and then comparing the phase relationship of these signals, the phase information at the antenna array is converted into amplitude differences with which it is possible to calculate the AOA of the signal. One such configuration to accomplish this is shown in Figure 8.10 [9]. In this case the amplitude relationship between the mode 1 output is compared with the mode 0 output of a *Butler matrix*.

One of the most significant advantages of this type of system is its speed of operation. All of the necessary computations are performed in the hardware and are instantaneously available. In fact, the appellation monopulse is based on this instantaneous operation and is borrowed from radar systems, where the angle of arrival is measured on a single pulse.

The devices used to perform the necessary phase shifts are called hybrids. These components shift the phase of the inputs in one of several possible ways. In this circuit both 90° hybrids and 180° hybrids are used as shown in Figure 8.10. For an 180° hybrid typically the phase-shifted inputs are summed for one of the

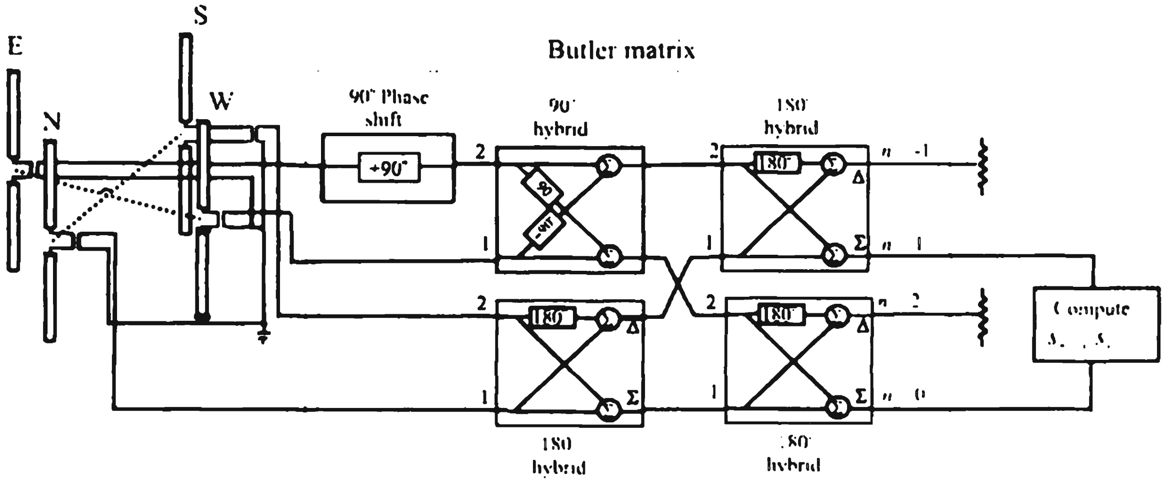


Figure 8.10 Monopulse direction finder using a Butler matrix for phase shifting.

outputs, while the other output sums one input with an 180°-shifted version of the other. The 90° hybrid shifts both inputs by –90° as shown.

By identifying the north antenna as antenna 0, west as antenna 1, south as antenna 2, and east as antenna 3 in the previous derivation then the operation of the monopulse direction finder with a Butler matrix can be analyzed as follows. By simply following the signals through the Butler matrix, the following results ensue.

$$s_{\mathrm{N}}(t) = s_0(t) = s(t)\cos\alpha e^{j\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi} \tag{8.49}$$

$$\begin{aligned} s_{\mathrm{W}}(t) = s_1(t) &= s(t)\cos\alpha e^{j\frac{2\pi R}{\lambda}\cos\alpha\cos(-\frac{2\pi}{4}+\varphi)} \\ &= s(t)\cos\alpha e^{j\frac{2\pi R}{\lambda}\cos\alpha[\cos(-\frac{\pi}{2})\cos\varphi-\sin(-\frac{\pi}{2})\sin\varphi]} \\ &= s(t)\cos\alpha e^{j\frac{2\pi R}{\lambda}\cos\alpha\sin\varphi} \end{aligned} \tag{8.50}$$

$$\begin{aligned} s_{\mathrm{E}}(t) = s_3(t) &= s(t)\cos\alpha e^{j\frac{2\pi R}{\lambda}\cos\alpha\cos(-\frac{6\pi}{4}+\varphi)} \\ &= s(t)\cos\alpha\times e^{j\frac{2\pi R}{\lambda}\cos\alpha[\cos(-\frac{3\pi}{2})\cos\varphi-\sin(-\frac{3\pi}{2})\sin\varphi]} \\ &= s(t)\cos\alpha e^{-j\frac{2\pi R}{\lambda}\cos\alpha\sin\varphi} \end{aligned} \tag{8.51}$$

$$\begin{aligned} s_{\mathrm{S}}(t) = s_2(t) &= s(t)\cos\alpha e^{j\frac{2\pi R}{\lambda}\cos\alpha\cos(-\pi+\varphi)} \\ &= s(t)\cos\alpha e^{j\frac{2\pi R}{\lambda}\cos\alpha[\cos(-\pi)\cos\varphi-\sin(-\pi)\sin\varphi]} \\ &= s(t)\cos\alpha e^{-j\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi} \end{aligned} \tag{8.52}$$

The mode zero output is therefore

$$\begin{aligned} s_{n=0}(t) &= s_{\mathrm{N}}(t) + s_{\mathrm{S}}(t) + s_{\mathrm{W}}(t) + s_{\mathrm{E}}(t)e^{j\frac{\pi}{2}}e^{-j\frac{\pi}{2}} \\ &= s(t)\cos\alpha e^{j\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi} + s(t)\cos\alpha e^{-j\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi} \\ &\quad + s(t)\cos\alpha e^{j\frac{2\pi R}{\lambda}\cos\alpha\sin\varphi} + s(t)\cos\alpha e^{-j\frac{2\pi R}{\lambda}\cos\alpha\sin\varphi} \end{aligned}$$

$$= s(t)\cos\alpha\left[\begin{array}{l} e^{j\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi} + e^{-j\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi} \\ \quad + e^{j\frac{2\pi R}{\lambda}\cos\alpha\sin\varphi} + e^{-j\frac{2\pi R}{\lambda}\cos\alpha\sin\varphi} \end{array}\right]$$

$$= s(t)\cos\alpha\left[2\cos\left(\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi\right) + 2\cos\left(\frac{2\pi R}{\lambda}\cos\alpha\sin\varphi\right)\right] \tag{8.53}$$

The term inside the brackets represents the phase shift induced on the signals from the antennas by the Butler matrix. The effects are perhaps most clearly seen by considering this term as a phasor, with amplitude given by

$$\text{Amplitude}_{n=0} = \sqrt{\left[2\cos\left(\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi\right) + 2\cos\left(\frac{2\pi R}{\lambda}\cos\alpha\sin\varphi\right)\right]^2}$$
$$= 2\cos\left(\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi\right) + 2\cos\left(\frac{2\pi R}{\lambda}\cos\alpha\sin\varphi\right) \tag{8.54}$$

Likewise, the $n = 1$ output of the Butler matrix is calculated as follows.

$$\begin{aligned} s_{n=1}(t) &= s_{\mathrm{N}}(t) + s_{\mathrm{S}}(t)e^{j\pi} + s_{\mathrm{W}}(t)e^{-j\frac{\pi}{2}} + s_{\mathrm{E}}(t)e^{j\frac{\pi}{2}} \\ &= s(t)\cos\alpha e^{j\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi} - s(t)\cos\alpha e^{-j\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi} \\ &\quad - js(t)\cos\alpha e^{j\frac{2\pi R}{\lambda}\cos\alpha\sin\varphi} + js(t)\cos\alpha e^{-j\frac{2\pi R}{\lambda}\cos\alpha\sin\varphi} \end{aligned} \tag{8.55}$$

The amplitude of $s_{n=1}$ is

$$\text{Amplitude}_{n=1} = \sqrt{4\sin^2\left(\frac{2\pi R}{\lambda}\cos\alpha\sin\varphi\right) + 4\sin^2\left(\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi\right)} \tag{8.56}$$

The amplitude response patterns for the monopulse array for $n = 0$ and $n = 1$ are shown in Figure 8.11.

In a similar fashion the output at the $n = -1$ port is calculated to be

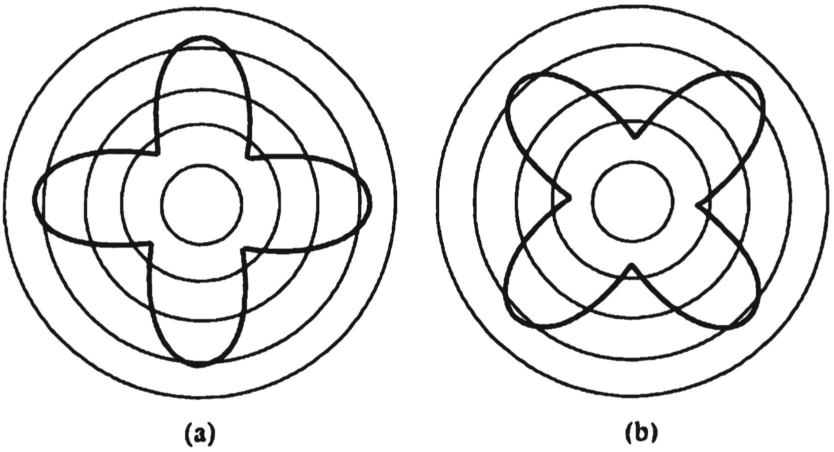


Figure 8.11 Amplitude response patterns for a monopulse array: (a) $n = 0$ and (b) $n = 1$.

$$
\begin{aligned}
s_{n=-1}(t) &= s(t)\cos\alpha\begin{bmatrix} e^{j\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi} - e^{-j\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi} \\ -je^{j\frac{2\pi R}{\lambda}\cos\alpha\sin\varphi} + je^{-j\frac{2\pi R}{\lambda}\cos\alpha\sin\varphi} \end{bmatrix} \\
&= s(t)\cos\alpha\begin{bmatrix} 2j\sin\left(\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi\right) \\ -\ 2j^{2}\sin\left(\frac{2\pi R}{\lambda}\cos\alpha\sin\varphi\right) \end{bmatrix} \\
&= s(t)\cos\alpha\begin{bmatrix} 2\sin\left(\frac{2\pi R}{\lambda}\cos\alpha\sin\varphi\right) \\ +\ j2\sin\left(\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi\right) \end{bmatrix} \\
&= s(t)\cos\alpha\begin{bmatrix} -2\sin\left(\frac{2\pi R}{\lambda}\cos\alpha\sin\varphi\right) \\ +\ j2\sin\left(\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi\right) \end{bmatrix}
\end{aligned}
\tag{8.57}
$$

with amplitude

$$
\text{Amplitude}_{n=-1} = \sqrt{4\sin^2\left(\frac{2\pi R}{\lambda}\cos\alpha\sin\varphi\right) + 4\sin^2\left(\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi\right)} \tag{8.58}
$$

Note that this amplitude is the same as the amplitude of the $n = 1$ port. It is the phase shifts that are different at these ports, not the amplitudes. Therefore, the amplitudes at either the $n = 1$ or the $n = -1$ port can be used to compare with the $n = 0$ port to determine the angle of arrival.

Finally, the signal at the $n = 2$ port is given by

$$s_{n=2}(t) = s(t)\cos\alpha\left[2\cos\left(\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi\right) - 2\cos\left(\frac{2\pi R}{\lambda}\cos\alpha\sin\varphi\right)\right] \tag{8.59}$$

The amplitude is given by

$$\begin{aligned}\text{Amplitude}_{n=2} &= \sqrt{\left[2\cos\left(\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi\right) - 2\cos\left(\frac{2\pi R}{\lambda}\cos\alpha\sin\varphi\right)\right]^2}\\ &= 2\cos\left(\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi\right) - 2\cos\left(\frac{2\pi R}{\lambda}\cos\alpha\sin\varphi\right)\end{aligned} \tag{8.60}$$

Note that there is a phase progression around the array, as we consider one antenna compared with another π/2 radians from it. For $n = 1$ and $n = -1$, the phase shift is in increments of 90°. For $n = 2$, this phase shift is in increments of 180° [0, 180; 0, 360; and 0, 540].

By taking the ratio of the signal at the $n = 1$ port to that of the $n = 0$ port, we get

$$\frac{s_{n=1}}{s_{n=0}} = \frac{2s(t)\cos\alpha\left[\sin\left(\frac{2\pi R}{\lambda}\cos\alpha\sin\varphi\right) + j\sin\left(\frac{2\pi r}{\lambda}\cos\alpha\cos\varphi\right)\right]}{2s(t)\cos\alpha\left[\cos\left(\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi\right) + \cos\left(\frac{2\pi R}{\lambda}\cos\alpha\sin\varphi\right)\right]} \tag{8.61}$$

Separating this into the real and imaginary parts and canceling appropriate terms yields

$$\text{Amplitude}_{1/0} = \left\{ \left[\frac{\sin\left(\frac{2\pi R}{\lambda}\cos\alpha\sin\varphi\right)}{\cos\left(\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi\right) + \cos\left(\frac{2\pi R}{\lambda}\cos\alpha\sin\varphi\right)} \right]^2 + \left[\frac{\sin\left(\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi\right)}{\cos\left(\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi\right) + \sin\left(\frac{2\pi R}{\lambda}\cos\alpha\sin\varphi\right)} \right]^2 \right\}^{1/2} \tag{8.62}$$

This amplitude ratio function is plotted in Figure 8.12 for $R/\lambda = 0.067$. This ratio is clearly dependent on the AOA of the signal and will yield different amplitudes depending on this angle. Note, however, that there are ambiguities. The ratio will yield the same result for signals arriving at several different angles around the array. Thus, additional information is required to resolve these ambiguities. There are no ambiguities if the signals are assumed to arrive from only a limited set of directions—for example, from 0 to $\pi/4$. If this is not the case, these ambiguities are usually resolved by using a sense antenna that is located at the center of the array. It provides a reference phase difference of zero. Similar results ensue when taking the ratio of the signal at $n = -1$ to that at $n = 0$—in fact, they are negatives of each other.

8.2.4 Amplitude Direction Finder

The antenna response patterns of directional antennas can be used to measure the direction of arrival of signals [10]. It is also possible to combine both amplitude and phase for direction finding [11]. Consider the antenna system of two dipoles configured in an Adcock configuration shown in Figure 8.13 with the amplitude

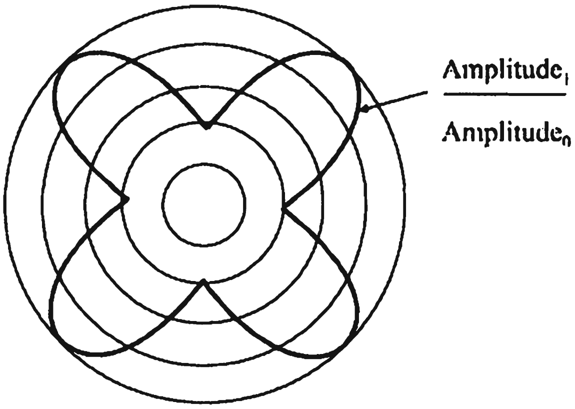


Figure 8.12 Amplitude ratio when $R = 0.2\text{m}$ and $f = 100$ MHz.

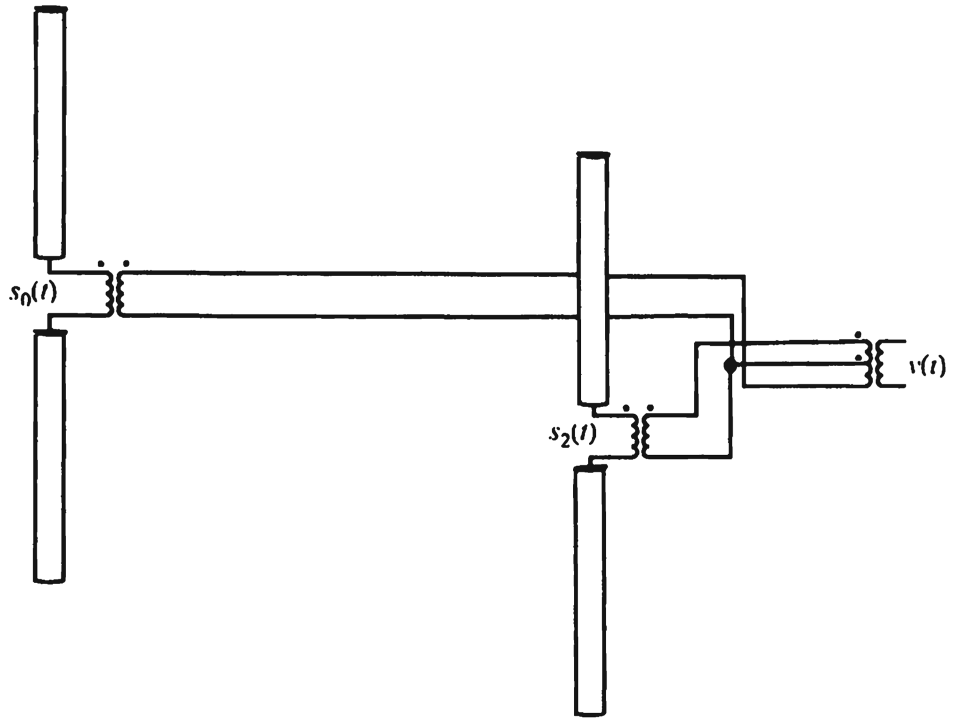


Figure 8.13 Adcock antenna array configuration.

response pattern shown in Figure 8.14. The amplitude of the signal arriving at the φ_1 shown would have a greater value than one arriving at φ_2 and that amplitude difference can be used as an indication of the AOA.

The output of this antenna configuration is given by

$$\begin{aligned}
v(t) &= s_0(t) - s_2(t) \\
&= s(t)\cos\alpha\left[e^{j\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi} - e^{-j\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi}\right] \\
&= s(t)\cos\alpha\left[\begin{array}{l}\cos\left(\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi\right)+j\sin\left(\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi\right) \\ \qquad -\cos\left(\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi\right)+j\sin\left(\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi\right)\end{array}\right] \\
&= s(t)\cos\alpha\left[j2\sin\left(\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi\right)\right] \qquad (8.63)
\end{aligned}$$

with amplitude

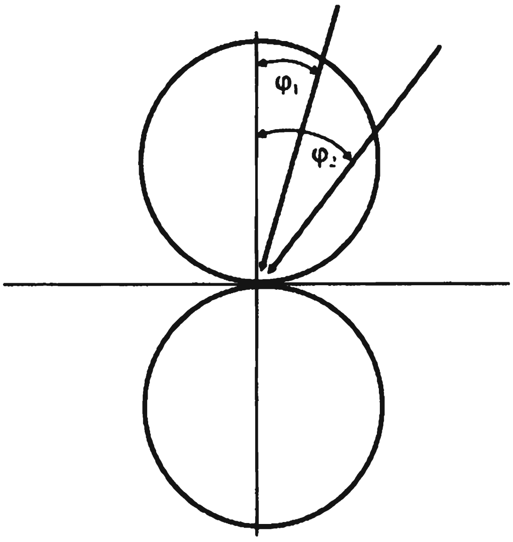


Figure 8.14 Amplitude response pattern for the Adcock antenna configuration shown in Figure 8.13.

$$\text{Amplitude}_{v(t)} = 2\sin\left(\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi\right) \tag{8.64}$$

Of course, in this simple case ambiguities can arise since there are other bearings that would produce the same amplitude. Another pair of dipoles arranged orthogonal to this antenna baseline could be incorporated to remove such ambiguities.

Since it is difficult to build antennas where the amplitude response is known exactly at all azimuths and all frequencies, especially in the lower-frequency ranges, amplitude comparison direction finding is only used when precise measurements of bearings are not important. Again, this antenna configuration is also a way to convert the phase of signals similar to the hybrids previously discussed.

8.2.4.1 Watson-Watt Direction Finder

This form of direction finder employs the amplitude measurement technique described above. It typically uses a circular array of dipoles, usually an orthogonally arranged set of the dipoles shown in Figure 8.15. The outputs of all four of these antennas are combined in phase to form a sense antenna, or a separate sense antenna is incorporated as shown.

Orthogonally arranged loop antennas can also be used in the Watson-Watt direction-finding architecture. The receiver arrangement is the same and, like the dipole arrangement, the orthogonal antenna's outputs can be combined to generate the sense output without a separate antenna.

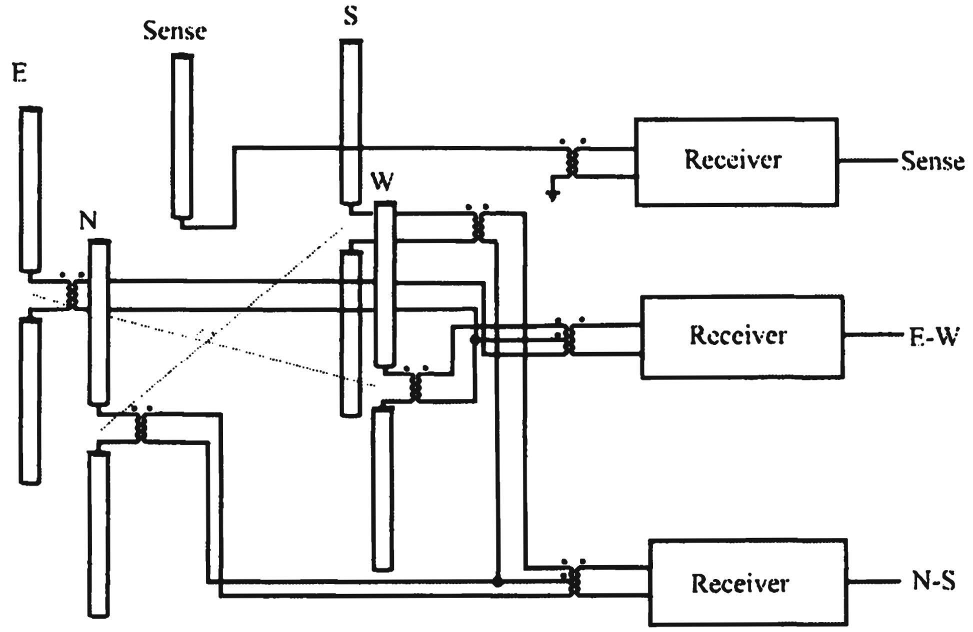


Figure 8.15 Watson-Watt direction-finding architecture.

In this arrangement, the AOA is determined as follows. The signal from the south antenna, $s_S(t)$, is subtracted from the signal from the north antenna, $s_N(t)$, yielding

$$
\begin{aligned}
s_{\mathrm{NS}}(t) &= s_{\mathrm{N}}(tt) - s_{\mathrm{S}}(t) = s(t)\cos\alpha\left[e^{j\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi} - e^{-j\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi}\right] \\
&= s(t)\cos\alpha\left[\begin{array}{l}
\cos\left(\dfrac{2\pi R}{\lambda}\cos\alpha\cos\varphi\right) \\
+\ j\sin\left(\dfrac{2\pi R}{\lambda}\cos\alpha\cos\varphi\right) \\
-\ \cos\left(\dfrac{2\pi R}{\lambda}\cos\alpha\cos\varphi\right) \\
+\ j\sin\left(\dfrac{2\pi R}{\lambda}\cos\alpha\cos\varphi\right)
\end{array}\right] \\
&= s(t)\cos\alpha\left[j2\sin\left(\frac{2\pi R}{\lambda}\cos\alpha\cos\varphi\right)\right] \qquad (8.65)
\end{aligned}
$$

Likewise, we subtract the signal from the west antenna, $s_W(t)$, from the signal at the east antenna, $s_E(t)$:

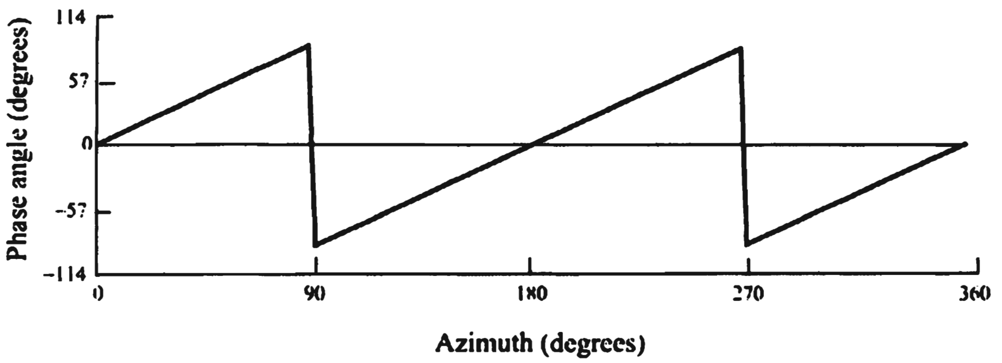


Figure 8.16 Indicated phase angle given by $\tan^{-1} s_{\text{EW}}(t)/s_{\text{NS}}(t)$.

$$
\begin{aligned}
s_{\text{EW}}(t) = s_{\text{E}}(t) - s_{\text{W}}(t) &= s(t)\cos\alpha\left[e^{j\frac{2\pi R}{\lambda}\cos\alpha\sin\varphi} - e^{-j\frac{2\pi R}{\lambda}\cos\alpha\sin\varphi}\right] \\
&= s(t)\cos\alpha\left[-j2\sin\left(\frac{2\pi R}{\lambda}\cos\alpha\sin\varphi\right)\right]
\end{aligned}
\tag{8.66}
$$

Next we form the ratio of the amplitude of these two quantities:

$$
\frac{s_{\text{EW}}(t)}{s_{\text{NS}}(t)} = \frac{-\sin\left(\dfrac{2\pi R}{\lambda}\cos\alpha\sin\varphi\right)}{\sin\left(\dfrac{2\pi R}{\lambda}\cos\alpha\cos\varphi\right)}
\tag{8.67}
$$

The Watson-Watt principle is based on approximating the AOA by the $\tan^{-1}$ of this ratio:

$$
\varphi \approx \tan^{-1}\frac{s_{\text{EW}}(t)}{s_{\text{NS}}(t)}
\tag{8.68}
$$

This ratio is plotted versus the azimuth in Figure 8.16 for R = 0.3m and f = 200 MHz.

The N-S and E-W antenna patterns are shown in Figure 8.17 for the values of R = 0.1m and f = 200 MHz (λ = 1.5m) yielding R/λ = 0.067, which are fairly typical parameters. The R/λ values are critical. Shown in Figure 8.18 are these patterns for several other values of R/λ.

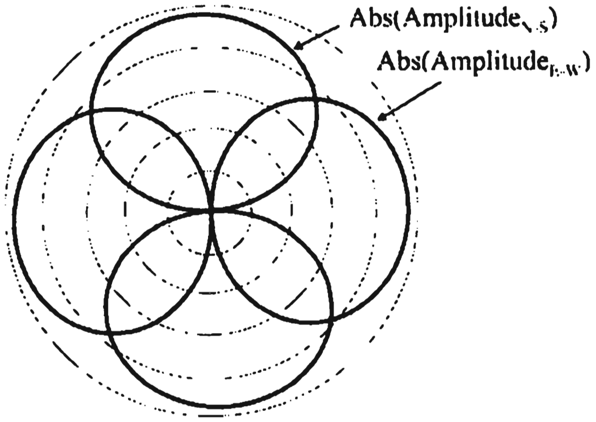
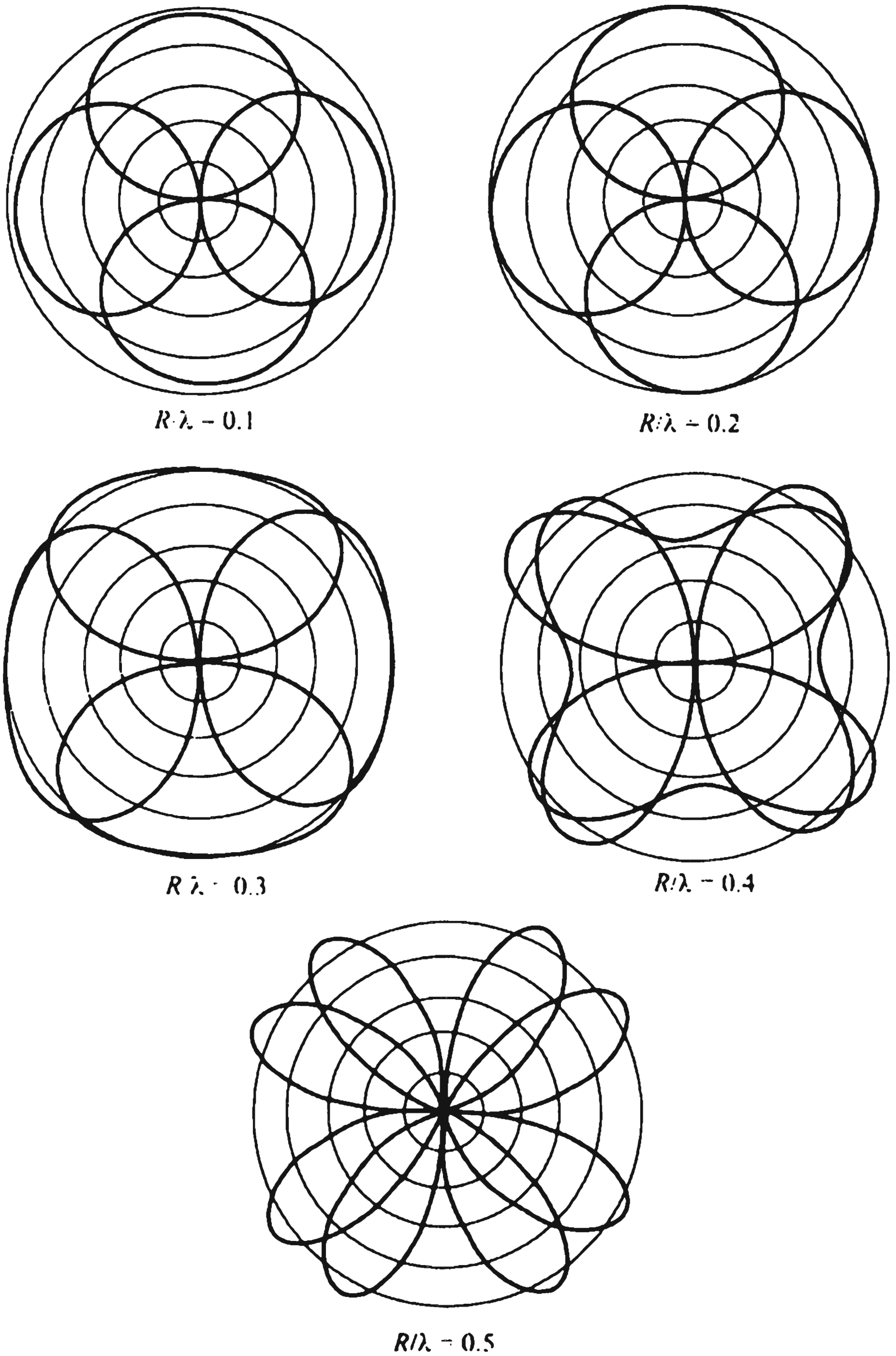


Figure 8.17 N-S and E-W antenna patterns for a Watson-Watt Adcock array. In this case R = 0.1m and f = 200 MHz. The absolute values are plotted for plotting convenience only.

The degree to which this approximation is not accurate is called the *spacing error*. Thus, plotted in Figure 8.19 is the true φ versus the function given by (8.59). For small values of R/λ this error is not too large, but it can get significant as R gets larger (or the wavelength gets shorter in relation to R). This characteristic is similar to the array response patterns discussed above. In fact, it is caused by the same effect. Note that the spacing error is a *systematic error* (caused by the characteristics of the system and is present all the time) and so can be removed by calibration in many cases. Such calibration usually involves sending signals toward the array from known directions, one at a time. The array response is measured and compared with the ground truth. The difference is stored in the calibration tables. Thus, plotted in Figure 8.19 is the true φ versus the function given by (8.59). For small values of R/λ this error is not too large, but it can become significant as R gets larger (or the wavelength gets shorter in relation to R). This characteristic is similar to the array response patterns discussed above. In fact, it is caused by the same effect. Note that the spacing error is a *systematic error* (caused by the characteristics of the system and is present all the time) and so can be removed by calibration in many cases. Such calibration usually involves sending signals toward the array from known directions, one at a time. The array response is measured and compared with the ground truth. The difference is stored in the calibration table. It may be necessary to do this at more than one elevation.

One of the advantages of this system is that it can be instantaneous. That is particularly important when trying to locate signals that have short duration. In order to be instantaneous, however, three receiver channels are necessary.



$R/\lambda = 0.1$

$R/\lambda = 0.2$

$R/\lambda = 0.3$

$R/\lambda = 0.4$

$R/\lambda = 0.5$

Figure 8.18 N-S and E-W response patterns for various values of R/λ.

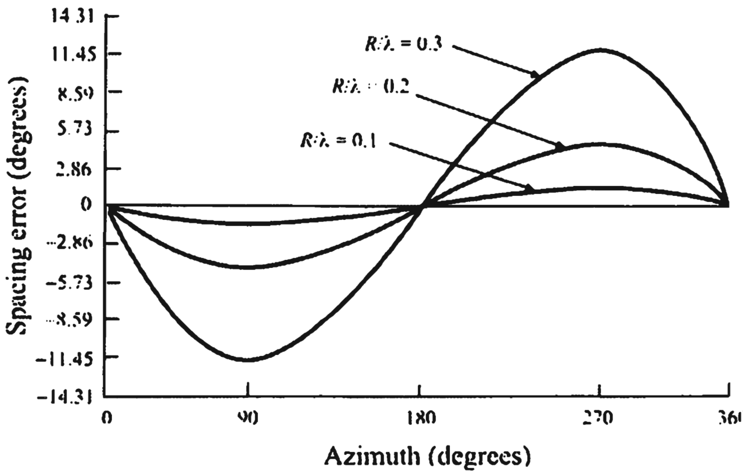


Figure 8.19 Spacing error caused by approximations to tan^{-1} $s_{E\text{-}W}/s_{N\text{-}S}$.

Alternately, if minimization of hardware is required, the antenna array patterns can be electronically scanned and a single receiver channel is adequate.

The antenna response patterns when combined with the sense data are shown in Figure 8.20. Dominant directional patterns evolve, yielding unambiguous bearings to be determined even from a single receiver. Also note that combining the signals, as here, is simply another way to accomplish the phase-shifting operation in the monopulse method.

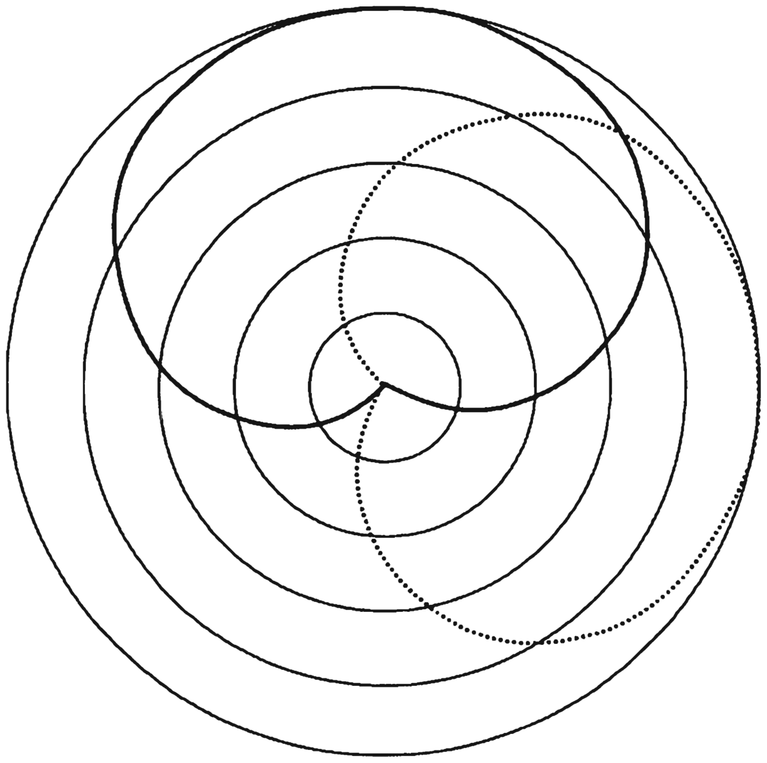


Figure 8.20 Antenna response patterns when the N-S and E-W patterns are combined with the sense antenna data.

8.2.5 Doppler Direction Finder

If an antenna is moving toward the source of a signal, then the frequency of the signal at the antenna is somewhat higher than if the antenna were still. If the antenna were moving away from the source of the signal, then the frequency would be somewhat lower than the true frequency. These frequency differences are an effect called *Doppler shift*, and are the same as when one listens to a passing train where, as the train is coming toward you, it sounds higher in frequency than after it passes and is moving away, when the audio signal is lower in frequency [12].

These effects can be exploited to determine arrival angles of signals as well. An antenna can be rotated in an EM field, as shown in Figure 8.21, and the output frequency measured. The frequency difference between the signal received at the rotating antenna and the true frequency measured at the stationary antenna determines the AOA.

Consider the diagram shown in Figure 8.22, which is a top view of Figure 8.21. The Doppler shift is given approximately by (assuming that $v << c$)

$$\Delta f = \frac{v}{c} f_0 \tag{8.69}$$

where v is the rotating antenna velocity in the direction of the signal, f_0 is the signal frequency, and c is the speed of propagation. The tangential velocity, v_t, shown in Figure 8.22 is given by

$$v_t = B\omega \tag{8.70}$$

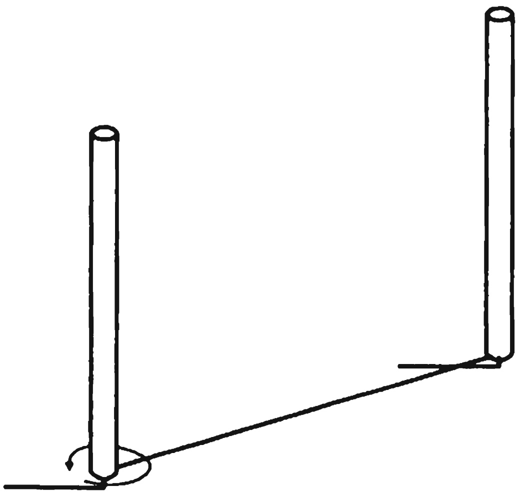


Figure 8.21 Doppler antenna configuration.

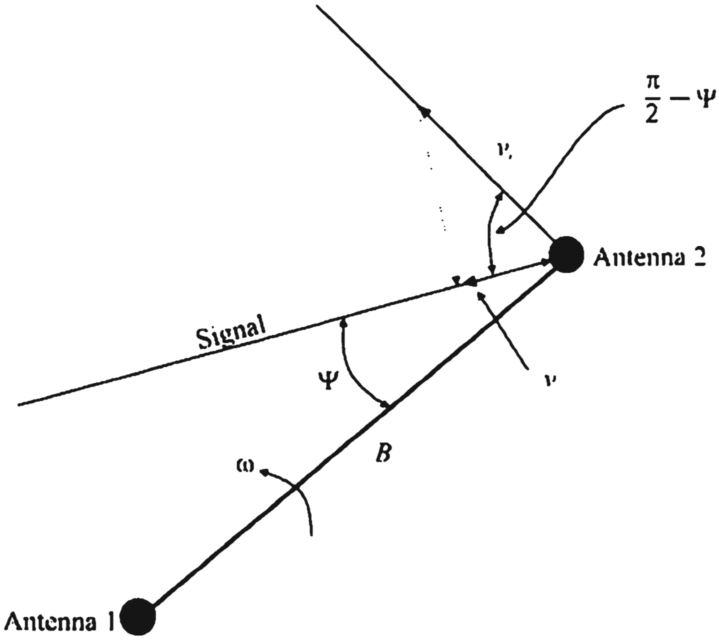


Figure 8.22 Top view of the Doppler rotating antenna array. Antenna 1 is stationary while antenna 2 rotates around it.

where B is the baseline length and ω is the angular velocity of the rotating antenna. The velocity v is related to the tangential velocity by

$$v = v_t \cos\left(\frac{\pi}{2} - \Psi\right) = v_t \sin \Psi \tag{8.71}$$

Therefore, the Doppler shift is given by

$$\Delta f = \frac{B\omega}{c} f_0 \sin \Psi \tag{8.72}$$

A sense antenna located in the center of the circle determines the true frequency. The response would be as shown in Figure 8.23. When the rotating antenna is moving toward the signal source, the difference frequency is positive,

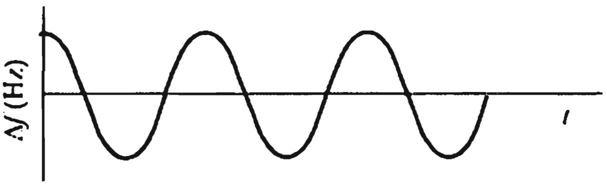


Figure 8.23 Output of the comparison of the two signals in a Doppler direction-finding system.

and when moving away, it is negative. In most implementations, a circular array of fixed antennas would be used as opposed to actually mechanically rotating an antenna. The effects of rotation can be achieved by electronically switching each antenna in the array in turn. The sense antenna is not really needed since the output of all the antennas in the circular array can be combined in phase to accomplish this function.

8.2.6 Array-Processing Bearing Estimation

In operational direction-finding applications, frequently there is more than one signal present at the frequency at which one wishes to determine an LOP. This is the case when two separate transmitters are using the same channel and it is also the case when one signal arrives at the antenna from two or more directions. The former case is called cochannel interference and the latter case is called multipath interference (see Chapter 4).

Many of the techniques described above do not correctly compute the bearing to any of the sources of the signals in these situations. Typically some other value is computed, sometimes being the vector sum of the signals, sometimes not, depending on the methods used. Therefore, some other way of finding the bearings is required.

The algorithms derived to deal with multiple signals are collectively called *array processing* direction finding, or sometimes called *superresolution* direction finding. Gething called this "multicomponent wave fields" [13]. There are three criteria that are normally associated with these algorithms. The first is their ability to resolve two signals that arrive at closely spaced azimuths. This is the criterion from which the term "superresolution" is derived. The closer in azimuth that two signals arrive, the more difficult it is to separate them—to calculate the two distinct AOAs. The second criterion is whether there is a statistical bias in the computed azimuths. This bias is reflected in incorrectly computing the azimuths, irrespective of how well they are resolved. Since the measurements of AOAs in real situations are statistical calculations, this bias is a bias of the mean of the density function away from the true values. These two criteria are often traded for one another, as increasing one can deleteriously affect the other. The third criterion is the variability, corresponding to the standard deviation of the probability density function, which represents the range of azimuths over which the calculations are expected to vary when noise is present.

Most of the high-resolution techniques require searching over the spaces spanned by the steering vectors. That is, it is assumed that signals are arriving from a particular direction and the steering vectors are calculated with that assumption. The correct answer is assumed to be the bearings where the largest spectrum results. Thus, for incremental azimuths around the antenna array, a set of linear equations are solved. The two techniques for calculating the azimuth angles of signals impinging on an antenna array discussed here are MUSIC and beamforming.

The *multiple-signal classification* (MUSIC) technique for spectral estimation is sensitive to fully coherent signals since then the covariance matrix is singular (described subsequently). Performance of the algorithm then depends on the degree of coherency in the signals. One advantage of MUSIC is that it can accommodate a nonuniform antenna array and no a priori knowledge of the number of signals is required.

There are various ways to compute the angular spectrum associated with an antenna and a particular algorithm, and each way has individual characteristics and limitations. These techniques vary in the assumptions about the signals. For example, it may be assumed that the two signals are not correlated with one another (correlated signals usually imply that one signal is a multipath reflection of the other). There are also assumptions about the configuration of the antenna array. One technique might require a circular array and another a linear array, for example. It should be noted that while this discussion talks about measuring the azimuth angle of arrival, the vertical angle of arrival could be calculated in the same way with an appropriately oriented antenna. This is particularly useful when computing the locations of HF targets, which could be arriving at a nonzero elevation angle, reflected off the ionosphere.

8.2.6.1 MUSIC

A method of superresolution direction finding based on eigen-decomposition of the signal correlation matrix has been devised, originally by Schmidt [14]; it is called MUSIC. In this technique the eigenvalues and eigenvectors of the signal correlation matrix are determined. The largest eigenvalues are assumed to be associated with signal vectors and the smallest eigenvalues are associated with noise vectors. Successive samples of antenna data are collected. Each of these samples corresponds to a single *frame*. Since, in general, RF signals are stochastic in nature, due to random noise as well as other reasons such as signal fading and modulation effects, the samples from frame to frame in general will be different.

Let μ_{jk} denote the contribution of signal k of *unit* amplitude to antenna j, and if the amplitude of signal k in frame i is ψ_{ki}, the signal ν_{jk} at antenna j without considering noise is determined by summing over all these signals. Thus, for S signals and N antennas

$$\begin{aligned}
\nu_{1i} &= \mu_{11}\psi_{1i} + \mu_{12}\psi_{2i} + \cdots + \mu_{1R}\psi_{Si} \\
\nu_{2i} &= \mu_{21}\psi_{1i} + \mu_{22}\psi_{2i} + \cdots + \mu_{2R}\psi_{Si} \\
&\vdots \\
\nu_{Ni} &= \mu_{N1}\psi_{1i} + \mu_{N2}\psi_{2i} + \cdots + \mu_{NR}\psi_{Si}
\end{aligned} \tag{8.73}$$

or

$$\nu_{ji} = \sum_{k=1}^{S} \mu_{jk}\psi_{ki} \tag{8.74}$$

In this expression, the product $\mu_{jk}\psi_{ki}$ represents the component of the signal at antenna j due to signal k during frame i. Note that if there is only one signal present, then there is only one term on the right side of these equations.

With noise components η_{ji} included, this becomes

$$\nu_{ki} = \sum_{k=1}^{S} \mu_{jk}\psi_{ki} + \eta_{ji} \tag{8.75}$$

that is,

$$\vec{\nu}_i = \mathbf{M}\vec{\varphi}_i + \vec{\eta}_i \tag{8.76}$$

In (8.66) $\vec{\nu}_i, \vec{\eta}_i$, and $\vec{\varphi}_i$ vary from frame to frame while matrix **M** is constant. **M** is a function of the constant antenna geometry and of the signal arrival time differences at the two antennas on a baseline (and thus the AOAs), assumed constant over the duration of interest. Again, if there is only one signal present, then **M** is a column vector. Such column vectors herein will be denoted by $\vec{m}^{<i>}$.

The columns of **M** are the steering vectors associated with the *i*th antenna and are given by

$$\vec{m}_i = e^{j2\pi\vec{k}\bullet\vec{r}_i} \tag{8.77}$$

This steering vector is the response of the *i*th antenna if a signal were coming from direction *k*.

$\vec{\eta}_i$ varies because of the random characteristics of Gaussian noise; $\vec{\varphi}_i$ varies too due to random variations in the signals caused by effects other than noise.

Such effects include fading and random modulation on the antenna signals used to derive $\bar{\varphi}_i$.

It is assumed that the noise is uncorrelated from one tap to the next as well as one frame to the next, and Gaussian, with zero mean and standard deviation σ. Using the notation $\mathcal{E}\{x\}$ for the expected value of x, we then have

$$\begin{aligned} \mathcal{E}\{\eta_{ji}\} &= 0 \\ \mathcal{E}\{\eta_{ji}\eta_{ki}\} &= \sigma^2\delta_{jk} \end{aligned} \tag{8.78}$$

where δ_{jk} is the Kronecker delta function. The expected value of a matrix is the matrix of the element expected values, namely,

$$\mathcal{E}\left\{\begin{bmatrix} x_1 & x_2 \\ x_3 & x_4 \end{bmatrix}\right\} = \begin{bmatrix} \mathcal{E}\{x_1\} & \mathcal{E}\{x_2\} \\ \mathcal{E}\{x_3\} & \mathcal{E}\{x_4\} \end{bmatrix} \tag{8.79}$$

The conjugate transpose of a matrix $\mathbf{G} = [g_{ij}]$ is that matrix denoted by $\mathbf{G}^{\mathrm{H}}$ with entries $g_{ij} = \bar{g}_{ji}$, where the overbar denotes conjugation. $\mathbf{G}^{\mathrm{H}}$ is sometimes called the *Hermitian* of $\mathbf{G}$. Also given matrix $\mathbf{G}$ such that $g_{ij} = \bar{g}_{ji}$, then $\mathbf{G}$ is said to be a Hermitian matrix.

The sample covariance matrix $\mathbf{\Gamma}$ is generated by forming matrix

$$\mathbf{Y} = \begin{bmatrix} \vec{\mathrm{v}}_1 & \vec{\mathrm{v}}_2 & \cdots & \vec{\mathrm{v}}_F \end{bmatrix} \tag{8.80}$$

where F is the number of frames collected. Each column of this matrix corresponds to a set of data samples from the antennas.

Then

$$\begin{aligned} \mathbf{\Gamma} &= \mathcal{E}\left\{\mathbf{YY}^{\mathrm{H}}\right\} \\ &= \mathcal{E}\left\{\begin{bmatrix} \vec{\mathrm{v}}_1 & \vec{\mathrm{v}}_2 & \cdots & \vec{\mathrm{v}}_F \end{bmatrix} \begin{bmatrix} \vec{\mathrm{v}}_1^{\mathrm{H}} \\ \vec{\mathrm{v}}_2^{\mathrm{H}} \\ \vdots \\ \vec{\mathrm{v}}_F^{\mathrm{H}} \end{bmatrix}\right\} \\ &= \mathcal{E}\left\{\left[\vec{\mathrm{v}}_1\vec{\mathrm{v}}_1^{\mathrm{H}} + \vec{\mathrm{v}}_2\vec{\mathrm{v}}_2^{\mathrm{H}} + \cdots + \vec{\mathrm{v}}_F\vec{\mathrm{v}}_F^{\mathrm{H}}\right]\right\} \\ &= \left[\mathcal{E}\left\{\vec{\mathrm{v}}_1\vec{\mathrm{v}}_1^{\mathrm{H}}\right\} + \mathcal{E}\left\{\vec{\mathrm{v}}_2\vec{\mathrm{v}}_2^{\mathrm{H}}\right\} + \cdots + \mathcal{E}\left\{\vec{\mathrm{v}}_F\vec{\mathrm{v}}_F^{\mathrm{H}}\right\}\right] \end{aligned} \tag{8.81}$$

Consider an arbitrary element of this matrix.

$$\begin{aligned}
\mathcal{E}\{\vec{v}_\iota \vec{v}_\iota^{\mathrm{H}}\} &= \mathcal{E}\{(\mathbf{M}\vec{\varphi}_\iota + \vec{\eta}_\iota)(\mathbf{M}\vec{v}_\iota + \vec{\eta}_\iota)^{\mathrm{H}}\} \\
&= \mathcal{E}\{(\mathbf{M}\vec{\varphi}_\iota + \vec{\eta}_\iota)(\vec{\varphi}_\iota^{\mathrm{H}}\mathbf{M}^{\mathrm{H}} + \vec{\eta}_\iota^{\mathrm{H}})\} \\
&= \mathcal{E}\{\mathbf{M}\vec{\varphi}_\iota\vec{\varphi}_\iota^{\mathrm{H}}\mathbf{M}^{\mathrm{H}} + \mathbf{M}\vec{\varphi}_\iota\vec{\eta}_\iota^{\mathrm{H}} + \vec{\eta}_\iota\vec{\varphi}_\iota^{\mathrm{H}}\mathbf{M}^{\mathrm{H}} + \vec{\eta}_\iota\vec{\eta}_\iota^{\mathrm{H}}\} \\
&= \mathcal{E}\{\mathbf{M}\vec{\varphi}_\iota\vec{\varphi}_\iota^{\mathrm{H}}\mathbf{M}^{\mathrm{H}}\} + \mathcal{E}\{\mathbf{M}\vec{\varphi}_\iota\vec{\eta}_\iota^{\mathrm{H}}\} + \mathcal{E}\{\vec{\eta}_\iota\vec{\varphi}_\iota^{\mathrm{H}}\mathbf{M}^{\mathrm{H}}\} + \mathcal{E}\{\vec{\eta}_\iota\vec{\eta}_\iota^{\mathrm{H}}\} \\
&= \mathbf{M}\mathcal{E}\left\{\vec{\varphi}_\iota\vec{\varphi}_\iota^{\mathrm{H}}\right\}\mathbf{M}^{\mathrm{H}} + \mathbf{M}\mathcal{E}\{\vec{\varphi}_\iota\vec{\eta}_\iota^{\mathrm{H}}\} + \mathcal{E}\{\vec{\eta}_\iota\vec{\varphi}_\iota^{\mathrm{H}}\}\mathbf{M}^{\mathrm{H}} + \mathcal{E}\{\vec{\eta}_\iota\vec{\eta}_\iota^{\mathrm{H}}\} \\
&= \mathbf{M}\mathbf{H}\mathbf{M}^{\mathrm{H}} + \mathbf{\Phi} \qquad (8.82)
\end{aligned}$$

The last step follows because the signals are assumed to be uncorrelated with the noises.

In this expression

$$\mathbf{H} = \mathcal{E}\left\{\vec{\varphi}_\iota\vec{\varphi}_\rho^{\mathrm{H}}\right\} = \mathcal{E}\left\{\begin{bmatrix}\varphi_{1\iota}\\ \varphi_{2\iota}\\ \varphi_{3\iota}\\ \vdots\\ \varphi_{S\iota}\end{bmatrix}\begin{bmatrix}\bar{\varphi}_{1\rho} & \bar{\varphi}_{2\rho} & \bar{\varphi}_{3\rho}\cdots\bar{\varphi}_{S\rho}\end{bmatrix}\right\}$$

$$= \mathcal{E}\left\{\begin{bmatrix}
\varphi_{1\iota}\bar{\varphi}_{1\rho} & \varphi_{1\iota}\bar{\varphi}_{2\rho} & \varphi_{1\iota}\bar{\varphi}_{3\rho} & \cdots & \varphi_{1\iota}\bar{\varphi}_{S\rho} \\
\varphi_{2\iota}\bar{\varphi}_{1\rho} & \varphi_{2\iota}\bar{\varphi}_{2\rho} & \varphi_{2\iota}\bar{\varphi}_{3\rho} & \cdots & \varphi_{2\iota}\bar{\varphi}_{S\rho} \\
\varphi_{3\iota}\bar{\varphi}_{1\rho} & \varphi_{3\iota}\bar{\varphi}_{2\rho} & \varphi_{3\iota}\bar{\varphi}_{3\rho} & \cdots & \varphi_{3\iota}\bar{\varphi}_{S\rho} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\varphi_{S\iota}\bar{\varphi}_{1\rho} & \varphi_{S\iota}\bar{\varphi}_{2\rho} & \varphi_{S\iota}\bar{\varphi}_{3\rho} & \cdots & \varphi_{S\iota}\bar{\varphi}_{S\rho}
\end{bmatrix}\right\} \qquad (8.83)$$

and

$$\mathbf{\Phi} = \mathcal{E}\left\{\vec{\eta}_\iota\vec{\eta}_\rho^{\mathrm{H}}\right\} = \mathcal{E}\left\{\begin{bmatrix}\eta_{1\iota}\\ \eta_{2\iota}\\ \eta_{3\iota}\\ \vdots\\ \eta_{N_\iota}\end{bmatrix}\begin{bmatrix}\eta_{1\rho} & \eta_{2\rho} & \eta_{3\rho}\cdots\eta_{N\rho}\end{bmatrix}\right\}$$

$$= \mathcal{E}\left\{\begin{bmatrix} \eta_{1\iota}\bar{\eta}_{1\rho} & \eta_{1\iota}\bar{\eta}_{2\rho} & \eta_{1\iota}\bar{\eta}_{3\rho} & \cdots & \eta_{1\iota}\bar{\eta}_{N\rho} \\ \eta_{2\iota}\bar{\eta}_{1\rho} & \eta_{2\iota}\bar{\eta}_{2\rho} & \eta_{2\iota}\bar{\eta}_{3\rho} & \cdots & \eta_{2\iota}\bar{\eta}_{N\rho} \\ \eta_{3\iota}\bar{\eta}_{1\rho} & \eta_{3\iota}\bar{\eta}_{2\rho} & \eta_{3\iota}\bar{\eta}_{3\rho} & \cdots & \eta_{3\iota}\bar{\eta}_{N\rho} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \eta_{N\iota}\bar{\eta}_{1\rho} & \eta_{N\iota}\bar{\eta}_{2\rho} & \eta_{N\iota}\bar{\eta}_{3\rho} & \cdots & \eta_{N\iota}\bar{\eta}_{N\rho} \end{bmatrix}\right\} \tag{8.84}$$

$\mathbf{\Gamma}$, **H**, and $\mathbf{\Phi}$ are square matrices of order ($N \times N$), ($S \times S$), and ($N \times N$), respectively. **M** and $\mathbf{M}^{\mathrm{H}}$ are of order ($N \times S$) and ($S \times N$), respectively.

When there is no noise, measurement errors, or other random effects, then when R is less than ($N - 1$), the system must be overdetermined and one or more linear relationships will be satisfied by the measured elements of $\vec{\mathrm{v}}_i$ on each frame and $\mathbf{\Gamma}$ will be of reduced rank r. It is said to have *nullity* ($N - r$).

With the above noise model applied to a large sample of frames, the off-diagonal elements of $\mathbf{\Phi}$ approach zero and (8.72) assumes the asymptotic form

$$\mathbf{\Gamma} = \mathbf{MHM}^{\mathrm{H}} + \sigma^2\mathbf{I} \tag{8.85}$$

For completely uncorrelated noise, the off-diagonal elements of **H** also tend to zero. The off-directional elements of Γ contain the directional information about the arriving signals and do not tend toward zero, but the leading diagonal elements are real and tend toward equality; the physical interpretation is that all antennas are expected to sense approximately the same power over a long sampling period.

A common way of estimating Γ is to average the frame data, namely,

$$\begin{aligned} \hat{\mathbf{\Gamma}} &= \frac{1}{F}\sum_{i=1}^{F}\vec{\mathrm{v}}_i\vec{\mathrm{v}}_i^{\mathrm{H}} \\ &= \frac{1}{F}\sum_{i=1}^{F}\begin{bmatrix} \mathrm{v}_{1i}\mathrm{v}_{1i}^{\mathrm{H}} & \mathrm{v}_{1i}\mathrm{v}_{2i}^{\mathrm{H}} & \cdots & \mathrm{v}_{1i}\mathrm{v}_{Ni}^{\mathrm{H}} \\ \mathrm{v}_{2i}\mathrm{v}_{1i}^{\mathrm{H}} & \mathrm{v}_{2i}\mathrm{v}_{2i}^{\mathrm{H}} & \cdots & \mathrm{v}_{2i}\mathrm{v}_{Ni}^{\mathrm{H}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{v}_{Ni}\mathrm{v}_{Ni}^{\mathrm{H}} & \mathrm{v}_{Ni}\mathrm{v}_{2i}^{\mathrm{H}} & \cdots & \mathrm{v}_{Ni}\mathrm{v}_{Ni}^{\mathrm{H}} \end{bmatrix} \end{aligned} \tag{8.86}$$

For the above approach to work, the frame data must be samples of a zero-mean statistical process. The reason for this is that the TDOA information is in the off-diagonal entries in $\hat{\mathbf{\Gamma}}$. If the frame data is not based on a zero mean process, the off-diagonal elements of the frame covariance matrices $\vec{\mathrm{v}}_i\vec{\mathrm{v}}_i^{H}$ will not tend to zero with averaging. They thus contribute to the off-diagonal elements of $\hat{\mathbf{\Gamma}}$ causing erroneous results.

Calculating the TDOA

The goal of this analysis is to ascertain the TDOAs of multiple signals arriving at a pair of antennas. The above discussion was a prelude establishing the basics of the processes to do just this.

An *eigenvector* of a square matrix **G** ($N \times N$) is a nonzero column vector $\vec{e}_i$ that satisfies

$$\mathbf{G}\vec{e}_i = \lambda_i \vec{e}_i \tag{8.87}$$

for an associated scalar λ_i known as its *eigenvalue*. The eigenvalues of **G** are found from the *characteristic equation* of **G** given by

$$\det[\mathbf{G} - \lambda \mathbf{I}] = 0 \tag{8.88}$$

where **I** is the identity matrix. Matrix **G** has N eigenvalues, not all of which are necessarily distinct. The eigenvectors associated with any λ form an independent set of vectors and form a basis for the linear manifold associated with λ. This linear manifold is referred to as the *eigenvector manifold associated with* λ. Once the eigenvectors are found, the associated eigenvectors are determined by solving (8.77).

When the eigenvalues of $\hat{\mathbf{\Gamma}}$ are found, they fall into two groups. The first group is of larger values than the second group which, when there is no noise or measurement error, are equal to each other and are of a small value. The number of larger eigenvalues is a measure of the number of constituent waveforms present. In most cases of interest $\mathbf{\Gamma}$ is singular since normally $S < N - 1$. MUSIC is based on the *singular value decomposition* on the covariance matrix $\mathbf{\Gamma}$.

Singular Value Decomposition

Any $m \times n$ matrix **Y** can be decomposed into

$$\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{H}} = \begin{bmatrix}\mathbf{U}_{\mathrm{s}} & \mathbf{U}_{\mathrm{o}}\end{bmatrix}\begin{bmatrix}\mathbf{\Sigma}_{\mathrm{s}} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{V}_{\mathrm{s}}^{\mathrm{H}}\\ \mathbf{V}_{\mathrm{o}}^{\mathrm{H}}\end{bmatrix} \tag{8.89}$$

where **U** is an $m \times n$ orthogonal matrix, **V** is an $n \times m$ orthogonal matrix, and $\mathbf{\Sigma}_s$ is an $r \times r$ diagonal matrix with real, nonnegative elements σ_i, $i = 1, 2, \ldots, r = \min(m, n)$ arranged in descending order

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \sigma_{r+2} = \cdots = 0 \tag{8.90}$$

These σ_i are called the *singular values* of **Y**. The first r columns of **U** are the left singular vectors of **Y**, and the first r columns of **V** are the right singular vectors of **Y**. The structure of $\mathbf{\Sigma}$ is

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1 & & & & & & \\ & \sigma_2 & & & \mathbf{0} & & \\ & & \ddots & & & & \\ & & & \sigma_r & & & \\ & & & & 0 & & \\ & \mathbf{0} & & & & 0 & \\ & & & & & & \ddots \\ & & & & & & & 0 \end{bmatrix}$$

$$= \begin{cases} \begin{bmatrix} \mathbf{\Sigma}_s \\ \mathbf{0} \end{bmatrix}, & \text{if } m \geq n \\ \begin{bmatrix} \mathbf{\Sigma}_s & \mathbf{0} \end{bmatrix}, & \text{if } m < n \end{cases} \tag{8.91}$$

where r is the rank of **Y**. Thus, $\mathbf{U}_s$ consists of the left singular vectors associated with the nonzero singular values and $\mathbf{U}_o$ consists of the left singular vectors associated with the zero singular values. These $\mathbf{U}_s$ vectors span the subspace consisting of the vectors in **M**. The vectors in $\mathbf{U}_o$ span the orthogonal subspace so

$$\vec{m}^{<i>\mathrm{H}}\mathbf{U}_o = \mathbf{0} \tag{8.92}$$

Also, the singular vectors are normalized so that $\mathbf{U}^{\mathrm{H}}\mathbf{U} = \mathbf{I}$.

U is the $m \times m$ matrix of orthonormalized row eigenvectors of $\mathbf{YY}^{\mathrm{T}}$ while **V** is the $n \times n$ matrix of orthonormalized column eigenvectors of $\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$. The singular values of **Y** are defined as the nonnegative square roots of the eigenvalues of $\mathbf{YY}^{\mathrm{T}}$.

The singular value decomposition is perhaps most known for solutions to the least squares problem. For the linear system defined by the set of equations

$$\mathbf{A}\vec{x} = \vec{b} \tag{8.93}$$

where **A** is not square, but of dimensions $m \times n$, and $\vec{x}$ and $\vec{b}$ are vectors, then the least squares solution is that value of $\vec{x}$ where

$$\min_{\vec{x}} \left\| \mathbf{A}\vec{x} - \vec{b} \right\| \tag{8.94}$$

which is

$$\hat{\vec{x}} = \mathbf{V}\boldsymbol{\Sigma}^{0}\mathbf{U}^{\mathrm{T}}\vec{b} \tag{8.95}$$

Matrix $\boldsymbol{\Sigma}^{0}$ is

$$\boldsymbol{\Sigma}^{0} = \begin{bmatrix} 1/\sigma_1 & & & & 0 & 0 & \cdots & 0 \\ & \ddots & & & & & & \\ & & 1/\sigma_r & & & \vdots & & \vdots \\ & & & 0 & & & & \\ & & & & \ddots & & & \\ & & & & 0 & 0 & \cdots & 0 \end{bmatrix} \tag{8.96}$$

then matrix

$$\mathbf{A}^{0} = \mathbf{V}\boldsymbol{\Sigma}^{0}\mathbf{U}^{\mathrm{T}} \tag{8.97}$$

is known as the *pseudoinverse* of **A**.

Note that if the zeros are removed from $\boldsymbol{\Sigma}^{0}$, and **U** and **V** are appropriately reduced in size, rank(**A**) = r, and the solution reduces to

$$\vec{x} = \left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)^{-1}\mathbf{A}^{\mathrm{T}}\vec{b} \tag{8.98}$$

which, with $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{T}$

$$\vec{x} = \mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{U}^{\mathrm{T}}\vec{b} \tag{8.99}$$

Because $\boldsymbol{\Gamma}$ is positive definite and Hermitian, (8.79) can be expressed as

$$\boldsymbol{\Gamma} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^{\mathrm{H}} \tag{8.100}$$

where **U** is unitary.[1] Note that if a vector $\bar{x}$ is orthogonal to $\mathbf{M}^{\mathrm{H}}$, then it is an eigenvector of $\boldsymbol{\Gamma}$ with eigenvalue σ^2 because

$$\boldsymbol{\Gamma}\bar{x} = \mathbf{MHM}^{\mathrm{H}}\bar{x} + \sigma^2\bar{x} = \sigma^2\bar{x} \tag{8.101}$$

The eigenvector of $\boldsymbol{\Gamma}$ with eigenvalue σ^2 lies in $\mathcal{N}(\mathbf{M}^{\mathrm{H}})$, the null space of $\mathbf{M}^{\mathrm{H}}$. Thus, the smallest $(N - L)$ nonzero eigenvalues are

$$\lambda_{L+1} = \lambda_{L+2} = \cdots\lambda_N = \sigma^2 \tag{8.102}$$

It is therefore possible to partition the eigenvectors into noise eigenvectors (corresponding to the smallest eigenvalues) and signal eigenvectors (corresponding to the largest eigenvalues) and the covariance matrix $\boldsymbol{\Gamma}$ can be written as

$$\boldsymbol{\Gamma} = \mathbf{U}_{\mathrm{s}}\boldsymbol{\Sigma}_{\mathrm{s}}\mathbf{U}_{\mathrm{s}}^{\mathrm{H}} + \mathbf{U}_{\mathrm{n}}\boldsymbol{\Sigma}_{\mathrm{n}}\mathbf{U}_{\mathrm{n}}^{\mathrm{H}} \tag{8.103}$$

Letting $\mathcal{R}(\mathbf{X})$ denote the range space of **X**, the range of **Q** is the orthogonal complement to the range of **M**, because

$$\mathcal{R}(\mathbf{Q}) = \mathcal{N}(\mathbf{M}^{\mathrm{H}}) =^{\perp} \mathcal{R}(\mathbf{M}) \tag{8.104}$$

where $\mathcal{N}(\mathrm{A})$ is the null space of **A**. Therefore,

$$\mathcal{R}(\mathbf{U}_{\mathrm{s}}) = \mathcal{R}(\mathbf{M}) \tag{8.105}$$

$$\mathcal{R}(\mathbf{U}_{\mathrm{n}}) =^{\perp} \mathcal{R}(\mathbf{M}^{\mathrm{H}}) \tag{8.106}$$

$\mathcal{R}(\mathbf{U}_{\mathrm{s}})$ is called the *signal subspace* and $\mathcal{R}(\mathbf{U}_{\mathrm{n}})$ is called the *noise subspace*. The projection operators onto these signal and noise subspaces are defined as

$$\mathbf{P}_{\mathrm{M}} = \mathbf{MM}^{\diamond} = \mathbf{U}_{\mathrm{s}}\left(\mathbf{U}_{\mathrm{s}}^{\mathrm{H}}\mathbf{U}_{\mathrm{s}}\right)^{-1}\mathbf{U}_{\mathrm{s}}^{\mathrm{H}} = \mathbf{U}_{\mathrm{s}}\mathbf{U}_{\mathrm{s}}^{\mathrm{H}} \tag{8.107}$$

$$\mathbf{P}_{\mathrm{M}}^{\perp} = \mathbf{I} - \mathbf{MM}^{\diamond} = \mathbf{U}_{\mathrm{n}}\left(\mathbf{U}_{\mathrm{n}}^{\mathrm{H}}\mathbf{U}_{\mathrm{n}}\right)^{-1}\mathbf{U}_{\mathrm{n}}^{\mathrm{H}} = \mathbf{U}_{\mathrm{n}}\mathbf{U}_{\mathrm{n}}^{\mathrm{H}} \tag{8.108}$$

[1] An $n \times n$ matrix **U** is *unitary* if $\mathbf{UU}^{\mathrm{H}} = \mathbf{U}^{\mathrm{H}}\mathbf{U} = \mathbf{I}_n$.

Determining the TDOAs

The eigenstructure of the covariance matrix Γ can be analyzed to determine the TDOAs. The normal nonsingular nature of Γ can cause numerical problems in some cases, however. In this technique the N-dimensional space is partitioned into the aforementioned signal subspace and the orthogonal noise subspace. The computation approach, starting from the covariance matrix estimate $\hat{\Gamma}$, requires the determination of the eigenvalues and eigenvectors of this matrix. As previously mentioned, the eigenvalues form two sets, one consisting of larger eigenvalues than the other. Starting with $k = 0$, and going until $k = N$, where k is the estimated number of eigenvalues, a logical dividing point is selected for the number of signals present. The eigenvectors corresponding to the signal subspace eigenvalues in the larger set (larger in value, not cardinality), denoted $\mathbf{R}_s$, span the signal space while the eigenvectors corresponding to the smaller set, denoted $\mathbf{R}_n$, span the orthogonal noise subspace.

The matrix $\mathbf{R}_n$ of eigenvectors spanning the noise subspace is used to calculate the function

$$P_{\text{MUSIC}} = \frac{1}{\boldsymbol{\Omega}^{\text{H}}\mathbf{R}_{\text{n}}\mathbf{R}_{\text{n}}^{\text{H}}\boldsymbol{\Omega}} \tag{8.109}$$

since the denominator in this expression tends to zero due to orthogonality. The peaks in this function are determined, and the S highest peaks are an estimate of where the time differences occur. A simulated plot of P_{MUSIC} might look like that in Figure 8.24 [15]. In this example there are two signals impinging on the antenna array, both at 45° azimuth, while the elevation difference is 10°. The abscissas on this chart are the azimuth angle of arrival and elevation angle of arrival. The ordinate is the amplitude response of the array in dB. The sharp peaks are evident.

MUSIC is but one example in the technical field with the appellation superresolution array processing. Other types of this processing have similar peaks to the ones in Figure 8.24, and each algorithm has its own benefits and drawbacks [15]. The principal drawback for MUSIC is its inability to DF fully correlated (coherent) signals.

8.2.6.2 Beamforming

One of the oldest techniques for dealing with the co-channel problem is *beamforming*. This technique is not normally considered to be in the superresolution family since historically the techniques used did not separate two signals, but suppressed the antenna response in all but a given direction. In this

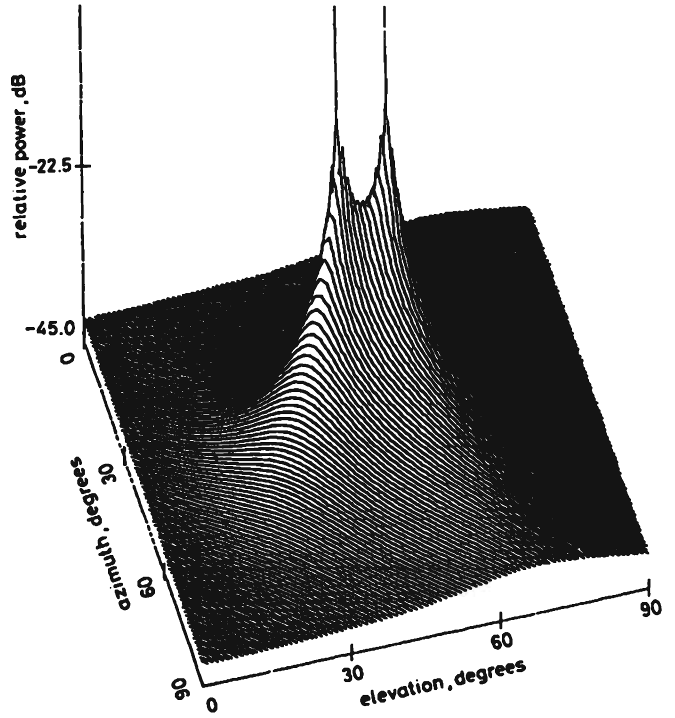


Figure 8.24 Simulated response of an array with signals at two azimuths and elevations: (1) Az = 45°, El = 40° and (2) Az = 45°, El = 50°. (*Source:* [15]. © HMSO 1978 and Peter Peregrinus, Ltd., 1990. Reprinted with permission.)

approach, the outputs of an array of K antennas are multiplied by weights, which are usually complex, and are delayed. The amplitude of the weight affects the amplitude of the signal from that particular antenna element while the phase of the weight changes the phase of the signal. The outputs of these multipliers and delay elements are then summed to form the array pattern. By properly adjusting the weights and time delays, the antenna's "beam" can be steered in the desired direction, enhancing the reception in that direction while suppressing the reception in others. The structure is shown in Figure 8.25.

If $x_j(t)$ represents the output of the jth antenna element and $s(t)$ is the signal that is transmitted then

$$x_j(t) = s(t + t_j) + n_j(t) \tag{8.110}$$

where $n_j(t)$ represents the additive noise at antenna element j. These signals are multiplied by weights with amplitudes w_j, delayed by a time τ_j, and then summed. The output of the beamformer is

$$y(t) = \sum_{j=1}^{K} w_j x_j(t - \tau_j) \tag{8.111}$$

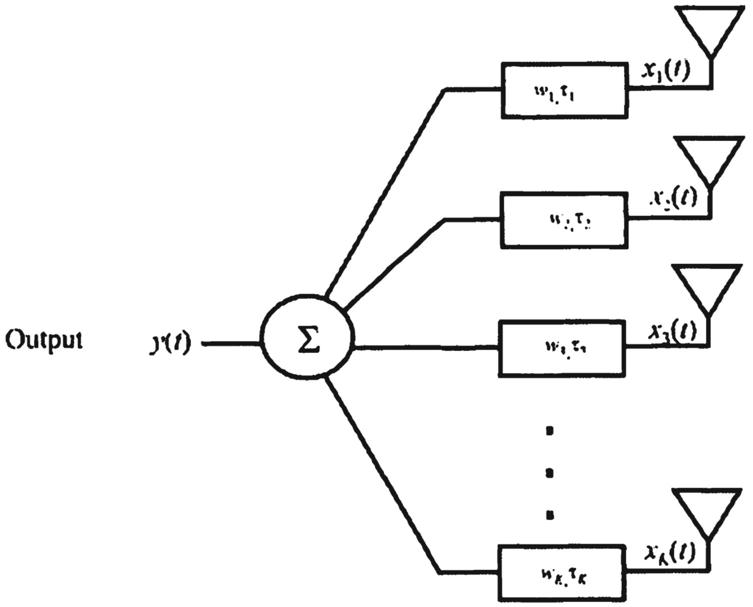


Figure 8.25 One of many types of beamforming network.

If the time delays are adjusted so that $t_j = \tau_j$, and assuming the weights are set to 1 then

$$\begin{aligned} y(t) &= Ks(t) + \sum_{j=1}^{K} n_j(t-\tau_j) \\ &= \sum_{j=1}^{K} w_j s(t + t_j - \tau_j) + w_j\, n_j(t-\tau_j) \end{aligned} \tag{8.112}$$

which under appropriate conditions will maximize $y(t)$. Search over a range of azimuths is accomplished by using several values of τ_j for each antenna path. Those directions where the response is maximum are assumed to be the direction of the arrival of signals.

Wullenwebber CDAA

One early form of beamformer direction finders was the Wullenwebber CDAA, a depiction of which is shown in Figure 8.26 [16]. Such direction finders are targeted against long-haul HF communications. They frequently use a goniometer, which is a mechanically rotated phase shifting networking typified by the network shown in Figure 8.27. The phase shifters form the CDAA pattern shown in Figure 8.28 and the direction where the null is formed is used to indicate the direction of

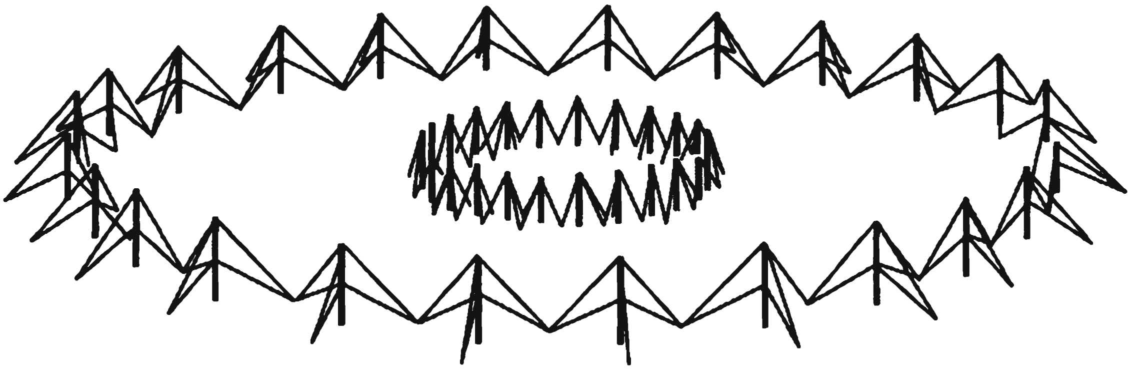


Figure 8.26 Wullenwebber CDAA. This antenna structure consists of two concentric antennas, each sized for different frequency bands.

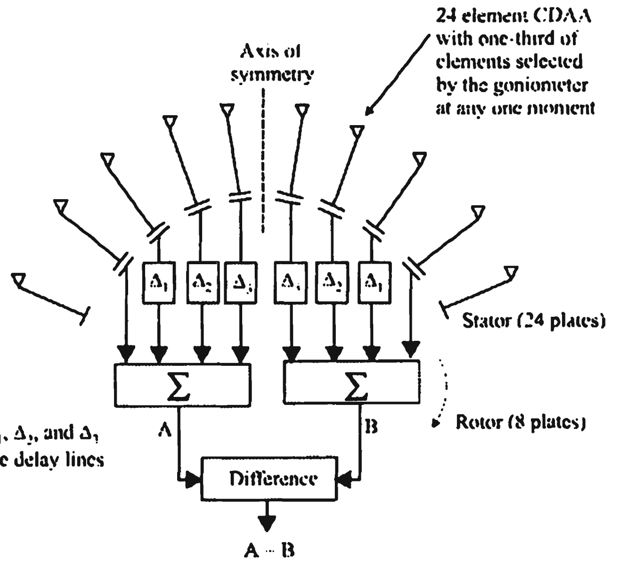


Figure 8.27 Phase difference network in a goniometer. (*Source:* [16]. © Controller HMSO 1978 and Peter Peregrinus, Ltd. 1990. Reprinted with permission.)

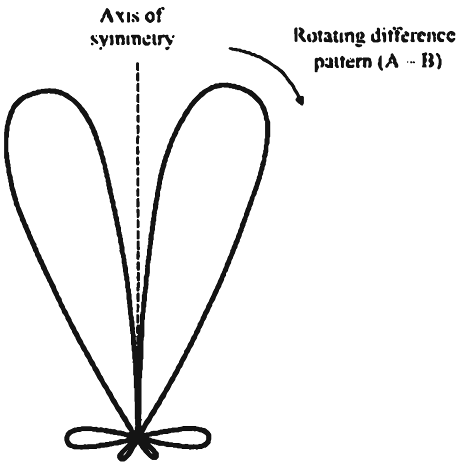


Figure 8.28 Wullenwebber antenna pattern with a goniometer. (*Source:* [16]. © Controller HMSO 1978 and Peter Peregrinus, Ltd. 1990. Reprinted with permission.)

arrival of HF signals [16]. Alternately, if the sum path of the phase-shifting network were used, a single maximum in the direction of arriving signals can be formed. The advantage of the null shown in Figure 8.28 is the beam edges are sharper where the measurement is made.

The Wullenwebber array is a manifestation of *wide-aperture direction finders* (WADF) [16]. Typically such antennas have two or more rings in order to cover the entire HF frequency range of 3 MHz (λ = 300m) to 30 MHz (λ = 30m). The larger array in Figure 8.26 has a typical diameter of 150m, while the higher frequency range inner ring has a diameter of typically 50m.

Care must be used at higher latitudes for such HF propagation paths, however, as the propagation modes are different from those at lower latitudes [17–19].

8.2.7 Line-of-Bearing (LOB) Optimization

The least squares range difference location problem has been investigated by Schmidt [20] and others. The range difference from a target to two sensors creates a hyperbolic line of position as shown in Figure 8.29. Thus, there are ambiguous answers for where the target is located. In order to ascertain where on this LOP the target lies, either one (or both) of the sensors must move to another location and another measurement taken, or a third sensor must be added creating a second (and unnecessary third) baseline. The resulting position fix determined by the intersection of the LOPs is then unique.

In the case shown, the sensors are widely separated. That need not be the case, however. The sensors could be two antennas in the same array, for example. The range difference in such an array is still a hyperboloid, but to calculate the target location this way is difficult. The LOPs intersect at too large a range from the

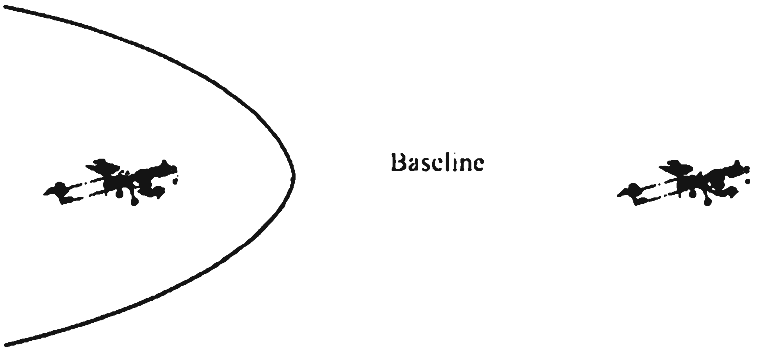


Figure 8.29 Two sensors creating a single baseline form a hyperbolic LOP upon which the target is located.

array. Small errors in calculating the LOPs result in large errors in calculating the target location because the LOPs are essentially parallel in the region that distance from the array.

Another technique to obtain a PF on a target is to measure the LOP from two or more sensors to the target. Where two or (preferably) more LOBs intersect is where the target is presumed to be located. One way of determining such lines of bearing is with phase interferometry where the phase difference is measured between $N > 2$ antennas and the LOB is computed as shown above in Section 8.2.2.

8.2.7.1 Phase Difference Averaging

Schmidt [20] showed that the TDOA averaging process produced the geolocation that was the closest feasible one in a least squared sense based on measured range differences. Although in those results the calculations produced geolocations based on range differences, herein they are somewhat extended to show that for an N-channel interferometer, calculating the signal phase differences in the same way produces bearings that are optimum in the least squares sense.

TDOA averaging is the process of removing the average TDOA, τ_{avg}, or equivalently the average range difference, Δ_{avg}, from the measured TDOAs or range differences. This produces feasible TDOAs or range differences that in a least squares sense are the closest to the measured data.

The top view of a triple channel interferometer is shown again in Figure 8.30. The phase angle of a planer signal crossing the array orthogonal to both the plane of the array and the antenna elements varies modulo 2π as the signal passes the array. Herein it is assumed that the baseline length of the array is small enough to avoid ambiguities at the highest operational frequency. The phase differences between the signals at the antennas taken a pair at a time are measured by one of several means not of consequence here. These phase angle differences are denoted

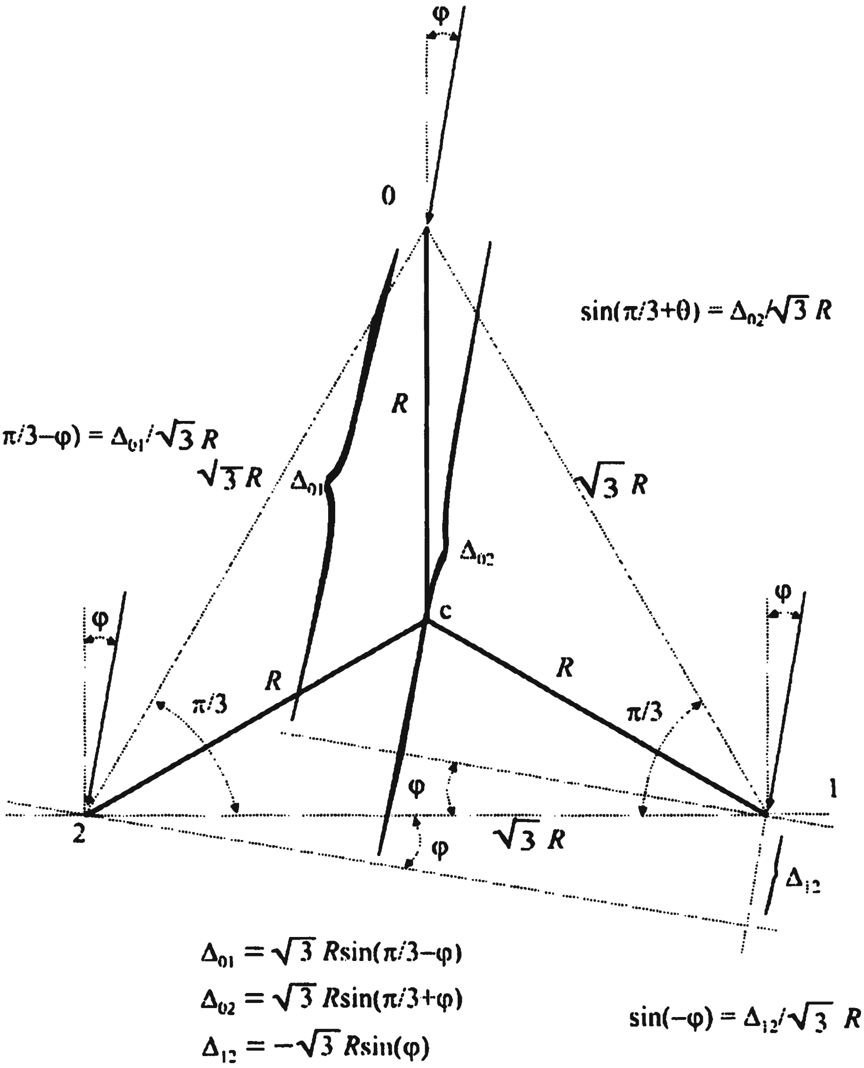


Figure 8.30 Top view of an $N = 3$ channel antenna array forming an interferometer.

ϕ_{ij}. They are related to the frequency and TDOAs at the antenna pairs τ_{ij} by $\phi_{ij} = 2\pi f \tau_{ij}$. Furthermore, since $\tau_{ij} = \Delta_{ij}/c$, where Δ_{ij} is the path distance traversed by the signal between antennas i and j and c is the speed of propagation of the wave in the medium, then $\phi_{ij} = 2\pi f\Delta_{ij}/c$. It is tacitly assumed that the sensor system is adequately far from the target that the signal is appropriately represented as a vertically polarized plane wave and that the AOA of the signal at each antenna is the same.

Ideally, the measured phase differences around any closed set of antennas sum to zero. In practice the sum will normally be nonzero due to measurement errors and noise. The process described by Schmidt [20] removes the measurement errors associated with measuring the range differences assuming that the sensors are widely spaced (i.e., not the antenna array shown in Figure 8.30, but separate sensors). The procedure described here assumes the sensors consist of the antenna array shown, with all antennas at the same geographic location, albeit separated somewhat by the baseline length shown.

In any case, the measured phase differences will not necessarily correspond to any feasible AOA due to the noise in the measurements. Therefore, as for the determination of the closest feasible location by Schmidt, the AOA based on the closest feasible phase angles is sought here. This will be called *phase difference averaging*. The resulting phase difference measurements are over-determined since there is more than one equation and only one dependent variable θ, thus yielding to least squares optimization.

The procedure described by Schmidt subtracts the average distances, or equivalently, the average TDOA, from each of the measurements, thereby calculating the closest feasible set of range differences. Thus,

$$\Delta_{\text{avg}} = \frac{1}{3}(\Delta_{01} + \Delta_{12} + \Delta_{20}) \tag{8.113}$$

and Δ_{avg} is subtracted from each measurement. Since $\Delta_{ij} = c\phi_{ij}/2\pi f$, then the average range difference can be expressed as

$$\Delta_{\text{avg}} = \frac{c}{2\pi f}\frac{1}{3}(\phi_{01} + \phi_{12} + \phi_{20}) \tag{8.114}$$

but

$$\frac{1}{3}(\phi_{01} + \phi_{12} + \phi_{20}) = \phi_{\text{avg}} \tag{8.115}$$

is the average of the measured phase differences. Therefore, subtracting the average phase difference from the measured phase differences will produce a set of adjusted phase differences that are feasible and least squares optimum.

As shown in Figure 8.31, the path-length differences for a signal with an AOA of φ are given by

$$\Delta_{01} = d_1 - d_0 = \sqrt{3}R\sin\left(\frac{\pi}{3} + \varphi\right)$$

$$\Delta_{02} = d_2 - d_0 = \sqrt{3}R\sin\left(\frac{\pi}{3} - \varphi\right) \tag{8.116}$$

$$\Delta_{12} = d_2 - d_1 = \sqrt{3}R\sin(-\varphi)$$

The relevant distances are Δ_{01}, Δ_{12} and $\Delta_{20} = -\Delta_{02}$. Care must be exercised when defining the signs of the angles involved. The d_i here represent the distance of antenna i from the target. Of course, over 2π radians these equations are ambiguous, each yielding two possible answers for the AOA. Any two taken at the same time, however, yield a unique answer.

The procedure for calculating the AOA is then:

1. Measure the phase differences between all pairs of antennas. This will yield $\binom{N}{2}$ phase difference measurements.
2. Determine the average path-length difference from (8.103).
3. Calculate the AOA φ using any two of (8.105).

This procedure works for any interferometric antenna array as long as the number of elements $N > 2$. When $N > 3$, the subarrays consisting of three elements each are taken one at a time and the procedure is applied to them.

8.3 Position-Fixing Algorithms

The previous discussions focused on calculating an LOP, upon which an emitting target is assumed to lie, based on measuring an azimuth and perhaps elevation AOAs of the EM wavefront at an antenna array. The target could lie at any point on the LOP, and therefore its location has not as yet been determined. Knowing just the LOP in some situations is useful—as an azimuth upon which to home a missile, for example. The direction in which to point a jammer antenna is another situation where just the LOP is useful information. It is generally not enough for most ES applications, however. Two or more such LOPs are usually combined to geolocate a target. If only two LOPs are available, the resultant fix calculated is called a *cut*. When three or more are used, it is called a *fix*.

A cut is less reliable than a fix. Two LOPs will always intersect at a single point, unless, of course, they are parallel. Three or more real LOPs will rarely intersect at the same point. Since determining the geolocation of a target is a statistical process, calculating the absolute location (as represented by a cut) is unreliable. Also, in general, for any statistical process the more relevant data that

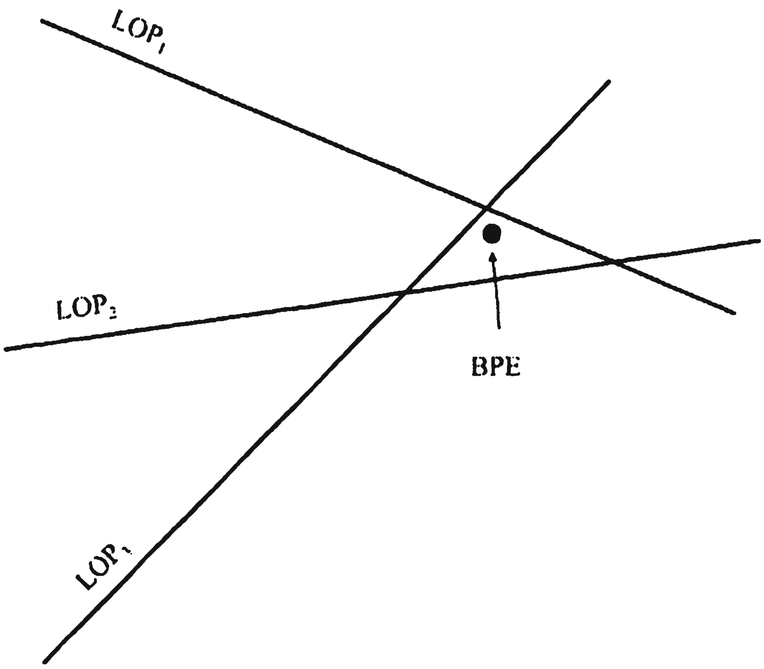


Figure 8.31 Determining the BPE. It will be some function of the individual LOPs.

can be brought to bear on the solution, the better because independent statistical results typically improve in accuracy inversely proportional to some power of the number of measurements made.

There are many algorithms available to facilitate the calculation of the location of an emitter based on multiple lines of position. This calculated location estimate is called the *best point estimate* (BPE). A few of these are presented in this section. The problem is illustrated in Figure 8.31. Normally, along with the BPE, a region is also calculated within which the emitter lies with a specified probability, usually 50% or 90%. The most common form of region is an ellipse, in which case it is called an EEP. Other forms are a circle, called a CEP, and a rectangle.

In these calculations, the a priori variance of the bearing for a system would typically be assumed to be the instrumental accuracy of the system, if the real accuracy is not known (by measurement, for example). This would typically be determined by experimentation and test, sometimes called *array calibration*, or just *calibration*. For ground-based systems, a turntable is typically used for this purpose. For larger aircraft, the aircraft are flown and data collected against the known location of emitters. For wide bandwidth systems, such calibration tables can become quite large.

8.3.1 Eliminating Wild Bearings

There is a technique for eliminating wild bearings when a sufficient number of bearings are available [21]. Such wild bearings, when included in BPE calculations, tend to yield erroneous results since by their very definition they do not point at the target. If it can be determined which bearings are "wild," they can be eliminated when the BPE is calculated.

Once the BPE has been calculated, determine the dispersion factor, given by

$$E = \sum \frac{\varphi_i^2}{\sigma_i^2} \tag{8.117}$$

where σ_i is the standard deviation (σ_i^2 is the variance) of the *i*th bearing and θ_i is the difference between the *i*th AOA and the *i*th line to the BPE. If there are wild bearings present, this number will be large. To eliminate these bearings, eliminate one bearing at a time and recalculate the dispersion factor. The best estimate will be the one associated with the smallest dispersion factor, and the bearing(s) excluded can be assumed wild.

A minimum of four LOPs are required in order to apply this technique. With three, eliminating one results in only two lines of position remaining; thus, a cut. It is impossible to tell which one of these BPEs is correct based on only two bearings. Therefore, a minimum of three LOPs at a time are required, necessitating at least four initially.

Care must be taken when eliminating bearings from the calculation of the BPE. The technique described here would apply if there were enough LOPs to start with, but it does not if there are not. In general, it is not known which bearings accurately correspond to those from the target and which might not. There is no statistical basis normally for eliminating bearings from the BPE calculation. There can be exceptions to this such as if the system operator "hears" interference while the bearing is being calculated. Another case would be if it is known that the system was inoperative when the bearing was measured.

8.3.2 Stansfield Fix Algorithm

One of the first algorithms developed for the purpose of calculating the location of an emitter based on multiple lines of bearing was due to Stansfield [22]. In that algorithm it is assumed that the bearing errors of the ES systems are normally (Gaussian) distributed. The joint probability density function of multiple lines of bearings is then a multivariate Gaussian probability density function. A maximum likelihood estimator for the BPE ensues by maximizing the exponent in the

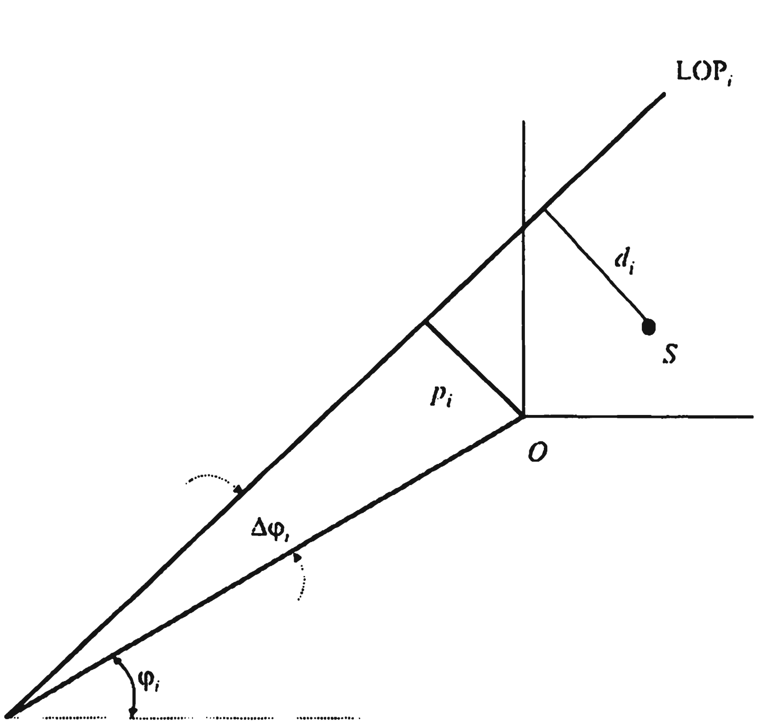


Figure 8.32 Definition of the terms for the derivation of the position fixing algorithm of Stansfield. (*Source:* [21]. © 1947, IRE. Reprinted with permission.)

equation for the joint probability density function (which will minimize the total probability of error because the exponent is negative).

Using Figure 8.32 to define the variables, then the joint probability of the miss distances between the LOPs and the true target location, denoted as p_i, is

$$P(p_1,\ p_2,\ \cdots\ ,\ p_N) = \frac{1}{(2\pi)^{N/2}\sum_{i=1}^{N}\sigma_{p_i}}\exp\left(-\frac{1}{2}\sum_{i=1}^{N}\frac{p_i^2}{\sigma_{p_i}^2}\right) \tag{8.118}$$

where N = number of LOPs.

It is assumed that the target lies at point $S = (x_\text{T}, y_\text{T})$, so that

$$d_i = p_i + x_\text{T}\sin\varphi_i - y_\text{T}\cos\varphi_i \tag{8.119}$$

where d_i is the distance from point S to the LOP$_i$. It is also assumed that the angular error is small so that $p_i \approx \Delta\varphi_i D_i$, and thus $\sigma_{pi} \approx D_i\sigma_\varphi$.

The joint probability of the d_i's is given by an expression similar to that above.

$$
\begin{aligned}
P(d_1, d_2, \cdots, d_N) &= \frac{1}{(2\pi)^{N/2}\sum_{i=1}^{N}\sigma_{p_i}}\exp\left(-\frac{1}{2}\sum_{i=1}^{N}\frac{d_i^2}{\sigma_{p_i}^2}\right) \\
&= \frac{1}{(2\pi)^{N/2}\sum_{i=1}^{N}\sigma_{p_i}}\exp\left[-\frac{1}{2}\sum_{i=1}^{N}\frac{(p_i + x_\mathrm{T}\sin\varphi_i - y_\mathrm{T}\cos\varphi_i)^2}{\sigma_{p_i}^2}\right]
\end{aligned}
\tag{8.120}
$$

The values of x_T and y_T that maximize the exponent in this expression are the coordinates with the highest probability of being those of the target—that is, the coordinates of the BPE. This yields the following expression for the target coordinates [22].

$$
\begin{aligned}
x_\mathrm{T} &= \frac{1}{ab - c^2}\sum_{i=1}^{N} p_i \frac{c\cos\varphi_i - b\sin\varphi_i}{\sigma_{p_i}^2} \\
y_\mathrm{T} &= \frac{1}{ab - c^2}\sum_{i=1}^{N} p_i \frac{a\cos\varphi_i - c\sin\kappa_i}{\sigma_{p_i}^2}
\end{aligned}
\tag{8.121}
$$

where

$$
\tan\varphi_i = \frac{y' - y_i}{x - x_i}
\qquad
d_i = a_i x_\mathrm{T} + b_i y_\mathrm{T} - c_i
$$

$$
\cos\varphi_i = \frac{d_i}{y' - y_\mathrm{T}}
\qquad
a = \sum_{i=1}^{N}\frac{\sin^2\varphi_i}{\sigma_{p_i}^2 D_i^2}
$$

$$
y' = (x_\mathrm{T} - x_i)\tan\varphi_i + y_i
\qquad
b = \sum_{i=1}^{N}\frac{\cos^2\varphi_i}{\sigma_{p_i}^2 D_i^2}
$$

$$
d_i = [(x_\mathrm{T} - x_i)\tan\varphi_i + y_i - y_\mathrm{T}]\cos\varphi_i
\qquad
c = \sum_{i=1}^{N}\frac{\sin\varphi_i\cos\varphi_i}{\sigma_{p_i}^2 D_i^2}
$$

and D_i is the distance between the true position and the system i.

8.3.3 Mean-Squared Distance Algorithm

An algorithm developed by Brown [23] based on an earlier algorithm by Legendre [24] will be presented in this section for calculating the BPE, which is based on minimizing the square of the miss distance of the BPE from the measured lines of position. Refer to Figure 8.33. The direction-finding systems are located at

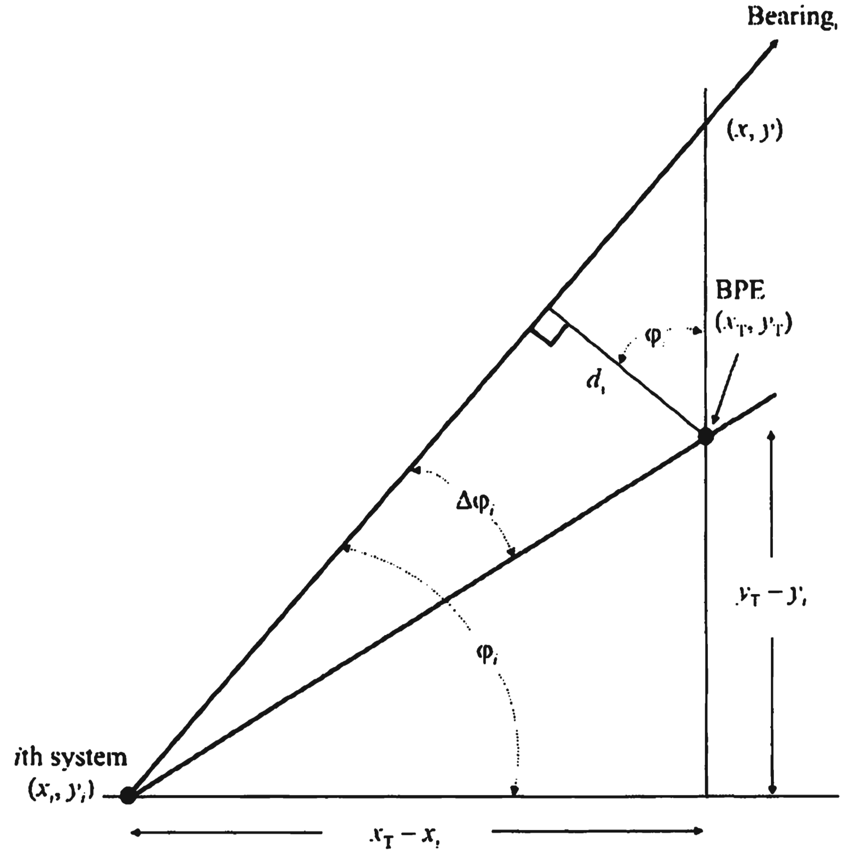


Figure 8.33 Definitions of the terms for derivation of Brown's mean-squared distance algorithm. (*Source:* [11].)

coordinates (x_i, y_i) and the bearing from the ith system is φ_i and is shown in Figure 8.33. The BPE is calculated to be (x_T, y_T) and the bearing error for the ith system is $\Delta\varphi_i$. To minimize the sum of the squares of the total miss distance, formulate

$$\begin{aligned} D &= \sum_{i=1}^{N} d_i^2 \\ &= \sum_{i=1}^{N} a_i^2 x_T^2 + \sum_{i=1}^{N} 2a_i b_i x_T y_T - \sum_{i=1}^{N} 2a_i c_i x_T \\ &\quad + \sum_{i=1}^{N} b_i^2 y_T^2 - \sum_{i=1}^{N} 2b_i c_i y_T + \sum_{i=1}^{N} c_i^2 \end{aligned} \tag{8.122}$$

where

$$a_i = \sin\varphi_i \qquad c_i = x_i \sin\varphi_i - y_i \cos\varphi_i$$

$$b_i = -\cos\varphi_i \qquad N = \text{number of LOPs}$$

Setting the first partial derivative of D with respect to x_T and then y_T equal to zero will find the values of x_T and y_T for which the total squared distance is minimized.

$$\begin{aligned} \frac{\partial D}{\partial x_\text{T}} &= 0 = 2x_\text{T}\sum_{i=1}^{N} a_i^2 + 2y_\text{T}\sum_{i=1}^{N} a_i b_i - 2\sum_{i=1}^{N} a_i c_i \\ \frac{\partial D}{\partial y_\text{T}} &= 0 = 2x_\text{T}\sum_{i=1}^{N} a_i b_i + 2y_\text{T}\sum_{i=1}^{N} b_i^2 - 2\sum_{i=1}^{N} b_i c_i \end{aligned} \tag{8.123}$$

which yield

$$x_\text{T} = \frac{\sum_{i=1}^{N} b_i^2 \sum_{i=1}^{N} a_i c_i - \sum_{i=1}^{N} a_i b_i \sum_{i=1}^{N} b_i c_i}{\sum_{i=1}^{N} a_i^2 \sum_{i=1}^{N} b_i^2 - \left(\sum_{i=1}^{N} a_i b_i\right)^2} \tag{8.124}$$

$$y_\text{T} = \frac{\sum_{i=1}^{N} a_i^2 \sum_{i=1}^{N} b_i c_i - \sum_{i=1}^{N} a_i b_i \sum_{i=1}^{N} a_i c_i}{\sum_{i=1}^{N} a_i^2 \sum_{i=1}^{N} b_i^2 - \left(\sum_{i=1}^{N} a_i b_i\right)^2} \tag{8.125}$$

Hertel extended these results using statistical estimation arguments (Hertel, R., personal communication, 1982) based on linear system theory. The above miss distance for system i is expressed as

$$d_i = a_i x_\text{T} + b_i y_\text{T} - c_i \tag{8.126}$$

where i is the ith measurement of a line of bearing and a_i, b_i, and c_i are as given above. In matrix form this is

$$\vec{d} = \mathbf{H}\vec{p} - \vec{c} \tag{8.127}$$

In this expression

$$\vec{c} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_N \end{bmatrix} \qquad \vec{p} = \begin{bmatrix} x_\text{T} \\ y_\text{T} \end{bmatrix} \qquad \mathbf{H} = \begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \\ a_3 & b_3 \\ \vdots & \vdots \\ a_N & b_N \end{bmatrix} \qquad \vec{d} = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ \vdots \\ d_N \end{bmatrix}$$

The least-squared error estimator for the target location vector $\vec{p}$ is given by [25]

$$\hat{\vec{p}} = \left[\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\right]^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\vec{c} \tag{8.128}$$

where, as usual, -1 denotes inverse and T denotes transpose. In this equation, **R** is a weighting matrix that is selected to optimize the calculation in some sense. The variance of this estimator is given by [25]

$$\mathbf{Q} = \left[\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\right]^{-1} = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix} \tag{8.129}$$

which is a covariance matrix [26]. The EEP parameters are related to the elements of this covariance matrix as follows:

$$\begin{aligned} L_{\mathrm{A}} &= \text{Semimajor axis} = \frac{2\left(\sigma_x^2\sigma_y^2 - \rho^2\sigma_x^2\sigma_y^2\right)C^2}{\sigma_x^2 + \sigma_y^2 - \left[\left(\sigma_y^2 - \sigma_x^2\right)^2 + 4\rho^2\sigma_x^2\sigma_y^2\right]^{1/2}} \\ L_{\mathrm{I}} &= \text{Semiminor axis} = \frac{2\left(\sigma_x^2\sigma_y^2 - \rho^2\sigma_x^2\sigma_y^2\right)C^2}{\sigma_x^2 + \sigma_y^2 + \left[\left(\sigma_y^2 - \sigma_x^2\right)^2 + 4\rho^2\sigma_x^2\sigma_y^2\right]^{1/2}} \end{aligned} \tag{8.130}$$

$$\tan 2\Psi = \frac{2\rho\sigma_x\sigma_y}{\sigma_y^2 - \sigma_x^2} \tag{8.131}$$

where

$$C = -2\ln(1 - P_e)$$

$$P_e = \text{Probability of being inside}$$

Here, Ψ is the tilt angle of the semimajor axis of the ellipse relative to the x-axis.

The weighting matrix $\mathbf{R}^{-1}$ in the above expression is used to optimize the performance. In one application of this algorithm $\mathbf{R}^{-1}$ is given by

$$\mathbf{R}^{-1}=\frac{1}{\sum_i \mathrm{QF}_i}\begin{bmatrix}\mathrm{QF}_1 & 0 & 0 & \cdots & 0\\ 0 & \mathrm{QF}_2 & 0 & \cdots & 0\\ 0 & 0 & \mathrm{QF}_3 & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \cdots & \mathrm{QF}_N\end{bmatrix}$$
$$\times\begin{bmatrix}\sigma_{d_1}^{-2} & 0 & 0 & \cdots & 0\\ 0 & \sigma_{d_2}^{-2} & 0 & \cdots & 0\\ 0 & 0 & \sigma_{d_3}^{-2} & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \cdots & \sigma_{d_N}^{-2}\end{bmatrix} \qquad (8.132)$$

where QF_i is some quality factor associated with measurement *i*. It might be the variance of the measurement, it could be some higher-order statistic as it makes sense, or it could be a measure of the SNR, as examples.

8.3.4 Combining Error Contours

The above algorithms for calculating the BPE of the location on an emitter also calculate an ellipse within which the emitter lies with a specified probability. If there are multiple BPEs, with associated EEPs, for a given emitter it is possible to combine these location estimates and their associated EEPs to refine the BPE and EEP. This is true assuming that the random variables involved have symmetric density functions, which includes the Gaussian case. An example of this is shown in Figure 8.33 where two original EEPs are combined. The resultant EEP is the smallest one shown in Figure 8.34. The original EEPs must be based on the same probability, for example, 50% and 90%. If this is not the case, adjustment of the original contours is necessary—fortunately such computations are simple, as given above. Blachman [27] derived the method for combining the EEPs based on two dimensions. These results were extended by Roecker [28] to include as many dimensions as desired using statistical estimation theory.

If the equations for the original EEPs are given by

$$a_i x^2 + b_i xy + c_i y^2 - d_i x - e_i y + f_i = 1 \qquad i = 1,2,\cdots,N \qquad (8.133)$$

with the centers of the ellipses denoted as α_i and β_i, the length of the semimajor axis denoted as L_{Ai}, the semiminor axis length denoted by L_{Ii}, and the angle of the semimajor axis with the *x*-axis given by Ψ_i, the parameters in this equation are:

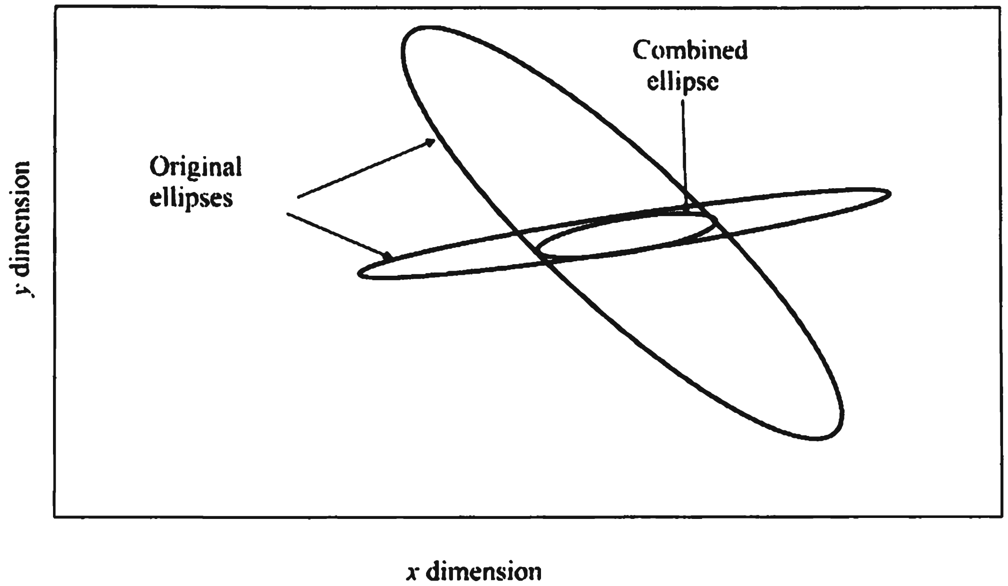


Figure 8.34 Combined error ellipses when they overlap.

$$
\begin{aligned}
a_i &= \frac{\cos^2 \Psi_i}{L_{Ai}^2} + \frac{\sin^2 \Psi_i}{L_{1i}^2} & \qquad b_i &= 2\cos\Psi_i \sin\Psi\left(\frac{1}{L_{Ai}^2} - \frac{1}{L_{1i}^2}\right) \\
d_i &= 2a_i\alpha_i + b_i\beta_i & & \\
e_i &= b_i\alpha_i + 2c_i\beta_i & \qquad c_i &= \frac{\sin^2 \Psi_i}{L_{Ai}^2} + \frac{\cos^2 \Psi_i}{L_{1i}^2} \\
f_i &= a_i\alpha_i^2 + b_i\alpha_i\beta_i + c_i\beta_i^2 & &
\end{aligned}
\tag{8.134}
$$

The parameters for the composite ellipse are determined from these parameters as follows. The equation for the composite ellipse is

$$ax^2 + bxy + cy^2 - dx - ey + f = 1 \tag{8.135}$$

with

$$
\begin{aligned}
a &= \sum_{i=1}^{N} a_i & \qquad e &= \sum_{i=1}^{N} e_i \\
b &= \sum_{i=1}^{N} b_i & \qquad \alpha &= \frac{2cd - ae}{4ac - b^2} \\
c &= \sum_{i=1}^{N} c_i & \qquad \beta &= \frac{2ae - bd}{4ac - b^2} \\
d &= \sum_{i=1}^{N} d_i & \qquad \Psi &= \frac{1}{2}\tan^{-1}\frac{-b}{c - a} \\
& & \qquad f &= a\alpha^2 + b\alpha\beta + c\beta^2
\end{aligned}
$$

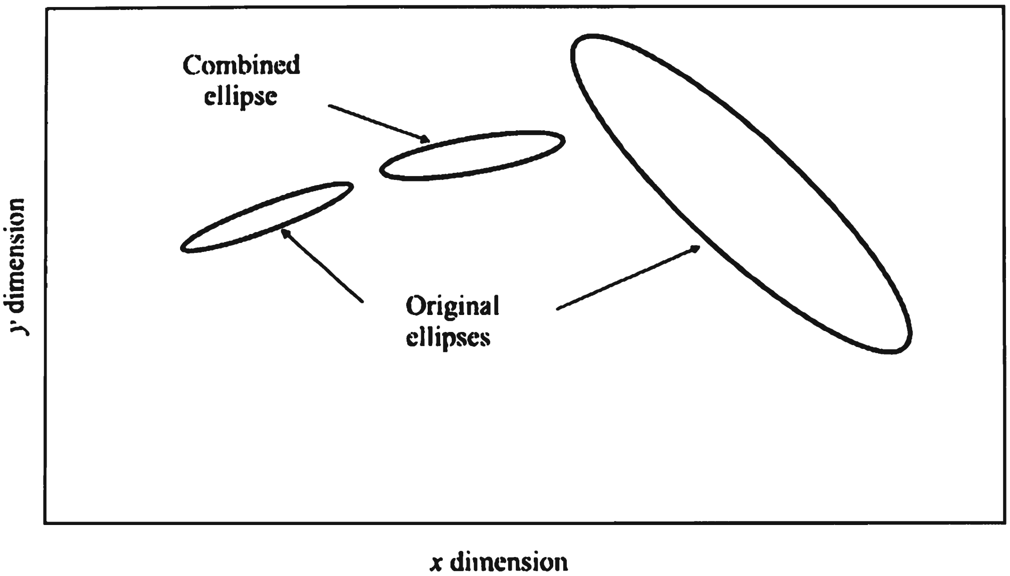


Figure 8.35 Combining error ellipses when the two original ellipses do not overlap can yield unexpected results as shown here.

The major and minor axes of the combined ellipse are given by

$$\begin{aligned}
L_A &= \sqrt{\frac{2}{a+c-\sqrt{(a-c)^2+b^2}}} \\
L_1 &= \sqrt{\frac{2}{a+c+\sqrt{(a-c)^2+b^2}}}
\end{aligned} \tag{8.136}$$

When the original ellipses overlap considerably as shown in Figure 8.34, the composite ellipse will be largely contained within the region of overlap of the original ellipses. If not, however, the combined ellipse can fall outside the confines of the original EEPs as shown in Figure 8.35. The combined ellipse, however, always has a smaller major and minor axis than the original EEPs. Therefore, as long as the above conditions are met, combining ellipses will always produce better (smaller) error contours for a given probability.

8.4 Single-Site Location Techniques

The source of HF signals propagating via skywave paths can be located by triangulation using any of the techniques discussed previously. However, it is also possible to use a single ES system for this purpose under some circumstances. If

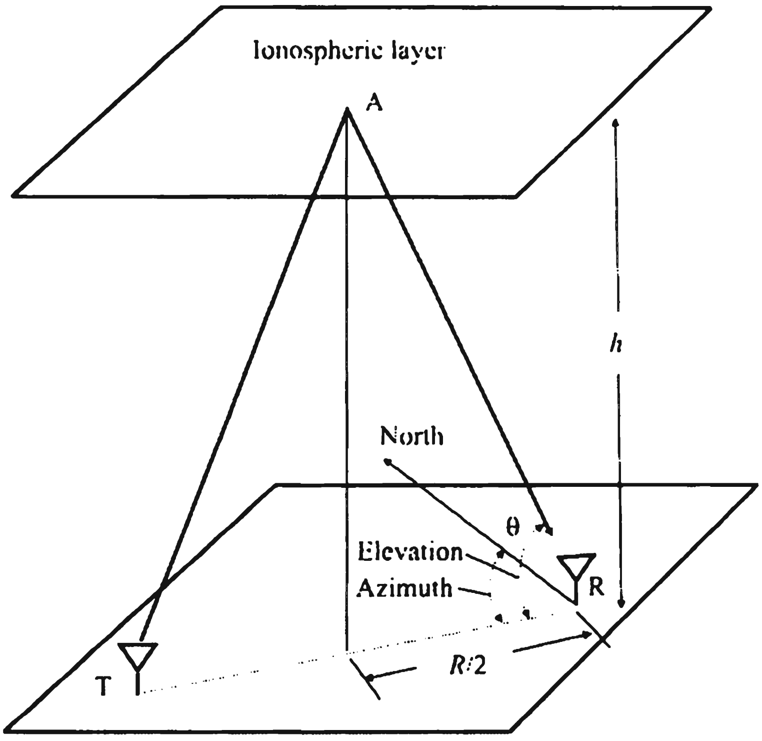


Figure 8.36 Single-site location of HF emitters.

the elevation of the ionospheric layer that the HF signal is refracting through, or, more accurately, the equivalent height of the layer that is reflecting the signal is known, as shown in Figure 8.36, then the emitter can be located since we know the range to the target and its angle of arrival. This technique takes advantage of the Breit-Tuve[2] and Martyn[3] equivalent path theorems in a flat ionosphere approximation [29]. The wave is assumed to be reflected at the midway point between the transmitter and the ES system. The elevation angle is related to the range and ionospheric height by

$$\tan\theta = \frac{h}{R/2} \tag{8.137}$$

so

[2] The Breit-Tuve theorem states that the group path length of a signal through the ionosphere is given by P' = TA + AR where TA is the distance between the transmitter (T) and the equivalent point of reflection in the ionosphere (A) and AR is the distance from A to the receiver (R). It assumes that the Earth is flat and that the ionosphere is horizontally stratified.

[3] Martyn's equivalent path theorem says that the virtual height of reflection in the ionosphere (*h* in Figure 8.36) is the same whether the signal is transmitted obliquely or vertically. This allows for vertical sounding at the receiver to be used to ascertain *h* at A (they are assumed to be the same).

$$R = \frac{2h}{\tan\theta} \tag{8.138}$$

Thus, by measuring the elevation AOA, the range to the target can be estimated. Although this derivation was given using planes for the ionosphere and the Earth, it can also be derived using spherical surfaces and is more accurate.

The devices used to measure the height of the ionosphere versus frequency are called *ionospheric sounders*. Typically a swept signal is radiated straight up for *vertical sounders* or at an angle for *oblique sounders*. The time of the return reflection is compared to when the signal was transmitted to ascertain the ionospheric height. The resultant display of the sounder results are called *ionograms*.

Vertical sounders, of course, measure the ionospheric height directly overhead. It is frequently assumed that the ionosphere is homogeneous, and the heights are the same throughout the region. The height is usually measured at the site of the ES system and it is assumed to be the same height at the reflection point. This is rarely true, so such techniques for measuring the location of transmitters are not that accurate. EEPs with axes that are 10% of the range to the target are typical. Thus, if the target is 100 km from the ES system, the major axis of the EEP can be 10 km or more. It should also be noted that although the reflecting surface discussed here was the ionosphere, any reflecting surface could also be used in some circumstances as long as the distance between that surface and the direction-finding system is known.

8.5 Fix Accuracy

Like any system that measures some physical phenomena, direction-finding systems always exhibit some degree of error. This error is typically both systematic and operational. The first represents errors that normally can, for the most part, be removed by calibration. The latter are caused by some characteristic of the environment and normally cannot be removed by calibration. These problems are solvable only by careful selection of the placement for the system.

The problem is to compute a fix on the emitter based on the LOPs computed at several separate physical locations. These LOPs could be obtained from different systems deployed on the ground or from an aircraft that is moving. Regardless, the situation is as depicted in Figure 8.31. The bearings do not normally cross at a point, but describe a region. Stansfield [22] called these regions "cocked hats."

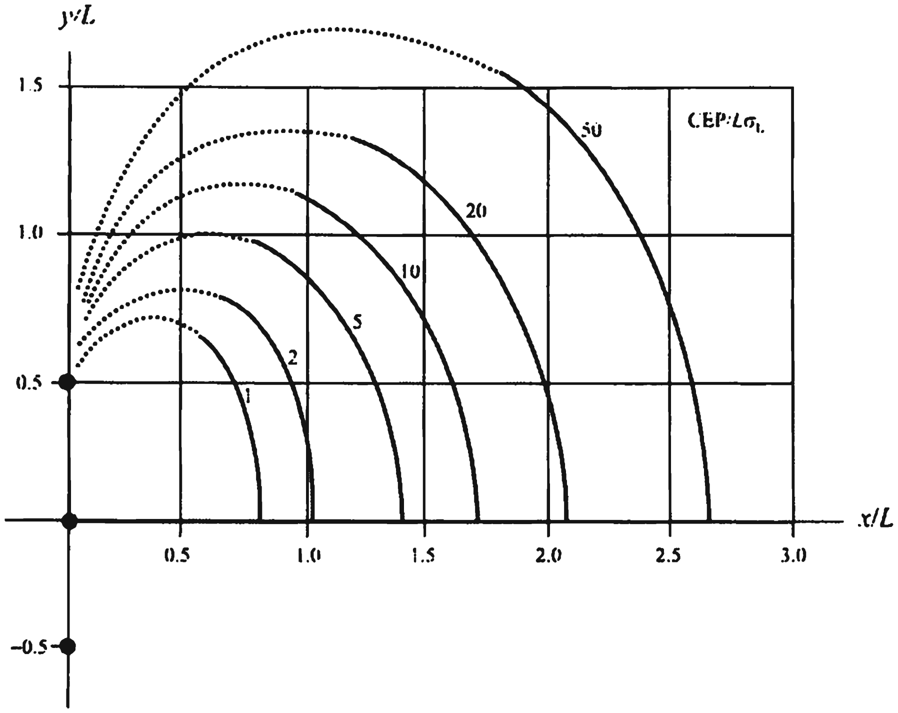


Figure 8.37 Contours of constant CEP/$L\sigma_L$ for a linear baseline. The sensors are indicated by •. (*Source:* [28]. © IEEE 1984. Reprinted with permission.)

8.5.1 Geometric Dilution of Precision

The configuration of the ES system baseline can affect the accuracy of the fixes computed by triangulation. This is referred to as *geometric dilution of precision* (GDOP). The effects can be seen by considering Figure 8.37 [30]. In this case a linear baseline is shown, with three ES systems placed at $y/L = -0.5$, 0, and 0.5, where L refers to the length of the baseline in consistent units. For this chart it is assumed that the ES systems are located close to the Earth's surface so that $1/R^4$ propagation characteristics predominate. Note that because of the baseline configuration, it is only necessary to show the first quadrant—the fourth quadrant is symmetric.

Where x/L is approximately 1.0 and y/L is approximately 0.5, the CEP/$L\sigma_L$ = 2, so if the baseline length L = 20 km and the system bearing standard deviation $\sigma = 5^\circ = 0.09$ radians, then the accuracy of the fix is

$$\text{CEP} = 2 \times 20 \text{ (km)} \times 0.09 \text{ (radians)} = 3{,}600\text{m}$$

As the target moves farther off the perpendicular bisector of the baseline, the accuracy degrades—the GDOP effect. Off the end of the baseline all of the

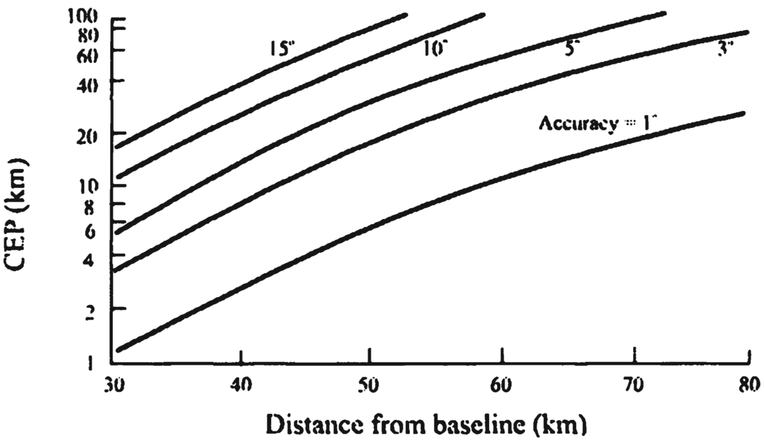


Figure 8.38 CEP versus range from the baseline on a perpendicular bisector.

bearings are the same and therefore do not intersect at a single point. In that case there is no solution and the CEP or EEP is unbounded.

The portion of these curves significantly off the perpendicular bisector represent regions where the CEP is not a very good measure of accuracy. In that case the two eigenvalues of the covariance matrix, λ_1 and λ_2 (for a two-dimensional PF calculation), are such that one is too large compared with the other, $\lambda_2/\lambda_1 < 0.01$. The calculation is changed to

$$\frac{L_A}{3.552\sqrt{\kappa}L\sigma} = \text{constant} \tag{8.139}$$

where

$$L_A = 2\sqrt{\kappa\lambda_1} \tag{8.140}$$

is the length of the major axis of the EEP and

$$\kappa = -2\ln(1 - P_e) \tag{8.141}$$

when P_e is the probability of being inside the EEP or CEP [30].

To illustrate the significance of these error bounds, a chart of the CEP of an emitter position fix when the emitter is located on the perpendicular bisector to the baseline is shown in Figure 8.38. In this case the baseline is linear as shown in Figure 8.37, and it is 30 km long from end to end. Shown is the 90% CEP as a function of the distance from the baseline and the accuracy of the direction-finding systems.

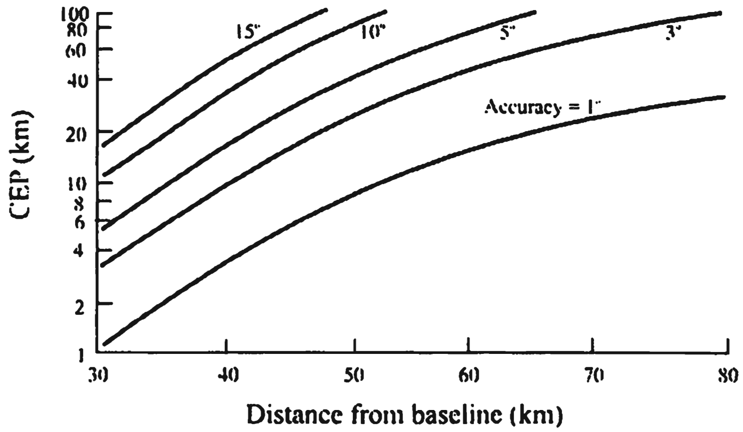


Figure 8.39 CEP versus range from the baseline on a 45° locus from the center of the baseline.

The effects of the GDOP can be seen in Figure 8.39, where a similar chart as that in Figure 8.38 is shown except that in this case the radial from the baseline is at a 45° angle from the center EW system. It is clear that off-axis, the accuracy is degraded.

A set of curves for a concave baseline is shown in Figure 8.40. The effects of baseline extensions are clear in this case as well. Extensions of the individual baselines made up by any pair of ES systems cause unbounded CEPs. Off the end of these baselines, even close fixes can have very large errors.

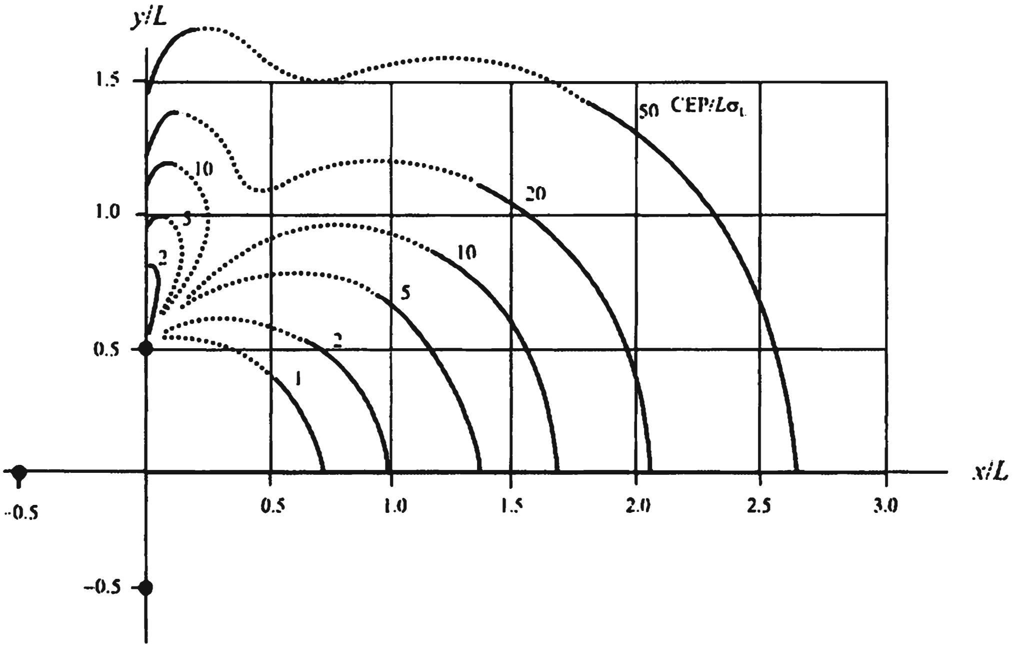


Figure 8.40 Contours of constant CEP/*Lσ* where the baseline is in the shape of a V. The sensors are indicated by •. (*Source*: [28]. © IEEE 1984. Reprinted with permission.)

The accuracy of the position fixes depends on the SNR at the ES system, as discussed previously. Elevating the ES system, therefore, normally will tend to improve the fix accuracy because the received power, given by (2.43), increases with antenna height while the noise tends to be the same or less. The noise can be less because some of it is caused by EMI generated by other electronic devices in the EW system. Raising the antenna moves it farther away from these devices.

8.6 Fix Coverage

Deployment of both ground-based and airborne EW systems is limited by several, but normally different, factors. Ground system deployment is largely dictated by the terrain available for deployment, with higher ground normally being desired so the coverage area is maximized. Airborne systems are limited by the weather as well as flight restrictions to avoid other air traffic.

High ground is not exclusively the province of EW systems. Other operational systems need such terrain as well, such as communication relays and ground surveillance radars. Therefore, planning on interference from other systems sharing the same frequency bands is wise.

It is important to keep in mind that, as opposed to determining the coverage area for signal intercept where the area is given by the disjunction of the coverage area of each individual system, the coverage area for geolocation is determined by the conjunction of the coverage area of two or more systems. The movement of one system to a different location can also achieve two or more system locations, of course. Thus, the geolocation coverage area of a suite of EW systems will always be less than the intercept coverage area.

Deployment of EW systems can influence the accuracy that can be achieved for locations of targets. The area to be covered can define better baselines for the bearing collection. In general, it is best for location accuracy purposes to keep the target region of interest along the perpendicular bisector of the pseudo-baseline which the EW systems make. This is obvious from Figures 8.37 and 8.40, which show the effects of GDOP. The best achievable accuracy is on this bisector, while the worst is off the ends of the pseudo-baseline.

When computing the BPE using LOPs, it is best to get the LOPs to cross at as close to a $\pi/4$ radian angle as possible. At angular crossings less than that, the targets are farther away than desirable. Equivalently, the baseline is shorter than desired. At angular crossings greater than that, the target is too close to the baseline and the GDOP will increase (in the limit the target lies on the pseudo-baseline and the accuracy goes to zero, which means the error goes to infinity). To achieve these angular specifications, an airborne system should not fly in a straight line if it is the only ES system being considered. It should fly in a semicircular pattern around the geographical region of interest as illustrated in Figure 8.41. The

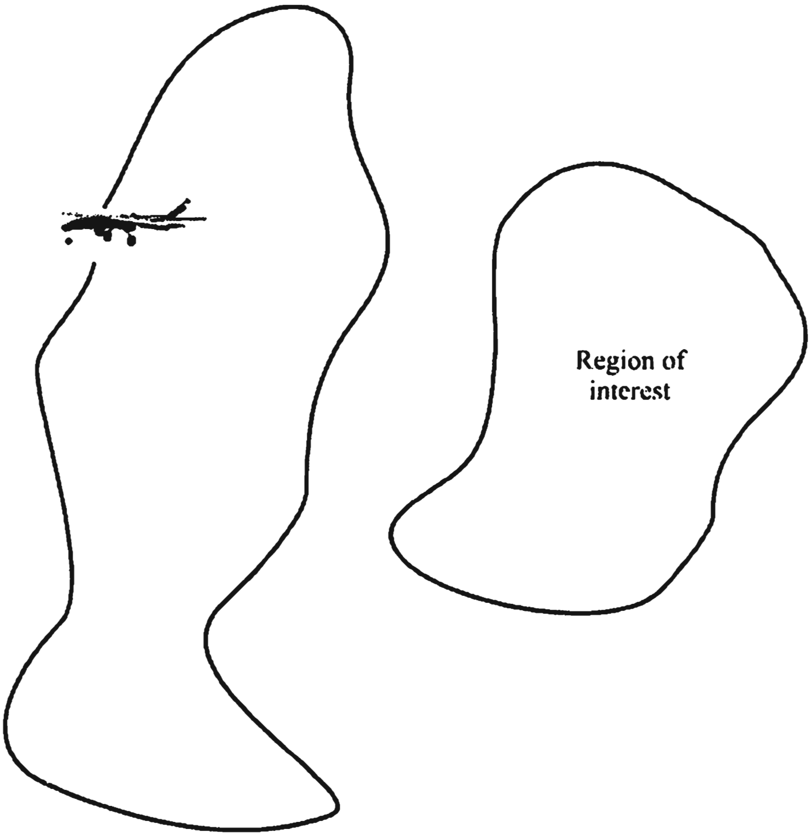


Figure 8.41 An aircraft flying a path that, as much as possible, surrounds the area of interest.

equivalent situation when ground systems are being considered is shown in Figure 8.42 where the baseline is shaped more or less in a semicircular arrangement.

For airborne systems that obviously are moving, a single platform can collect several bearings in time succession. If targets are emitting for enough time, then such approaches produce enough bearings to compute a fix. For frequency agile targets this may not be the case. When UASs are employed as EW sensor systems, then overflight of the target area is possible. This not only improves coverage of especially low-power targets, it also improves direction-finding geometry if employed correctly.

For ground-based situations, the collection systems are not necessarily moving when collecting data. Therefore, more than one EW system is typically necessary to collect bearings simultaneously with other systems. The most general configuration would consist of both ground systems and airborne systems operating simultaneously. (For ship applications at sea and line-of-sight EW, only the airborne case would normally apply since if EW is executed from a ship, the target is much closer than the ship commander would likely tolerate. For HF, these descriptions do apply since HF signals can propagate much farther.)

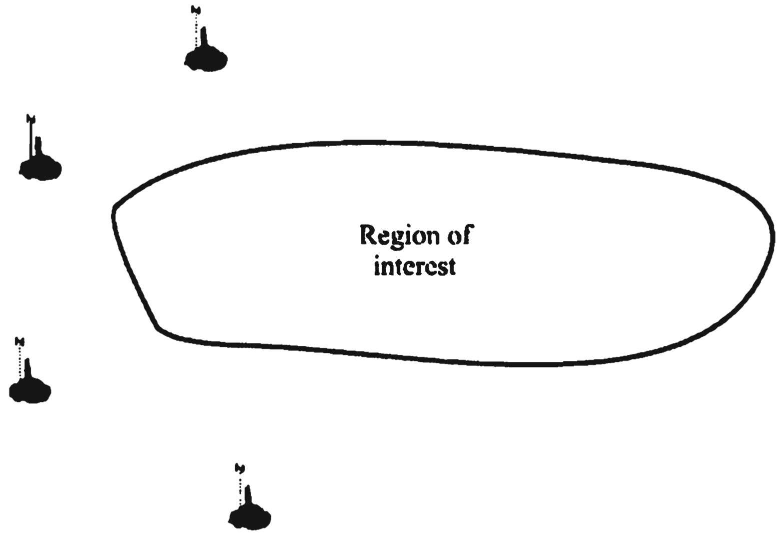


Figure 8.42 If the target area is narrow and deep, the EW systems should be configured in a concave configuration to maximize position fix accuracy.

Irrespective of whether the EW system is airborne or ground, if the area of interest is broad and shallow, then the best baseline configuration is one that is convex, as shown in Figure 8.43. This configuration provides better coverage from three of the four EW systems on each side while sacrificing the bearing from the fourth system for the width of coverage. In essence this configuration tries to place a straight baseline against the targets in one part of the coverage area.

On the other hand, if the area to be covered is narrow and deep, then a concave baseline is best as shown in Figure 8.41. Here all four bearings are used, sacrificing the width of coverage.

When there are only three ground systems available and the coverage area is nonlinear, then the quality of coverage is going to be somewhat less than that available with standoff sensor systems. The term *nonlinear* in this case refers to when friendly forces are deployed intermixed with both neutral population and enemy forces. Lacking any information about specific areas to be covered it can be shown that the best arrangement to cover the whole, more or less circular, area of interest is an equilateral triangle as shown in Figure 8.44. However, due to the GDOP effects described above, fix accuracy in line with any of the baselines formed with the systems will degrade as shown in the figure. Augmenting this arrangement with a direction-finding system on a UAS linked with the three ground systems will significantly improve the coverage accuracy as well as the coverage area.

Bordering on optimal direction-finding geometries can be obtained with two or more standoff aircraft, be they manned fixed wing, manned rotary winged, or UAS with UAS augmentation. The UAS in this case is assumed to be capable of

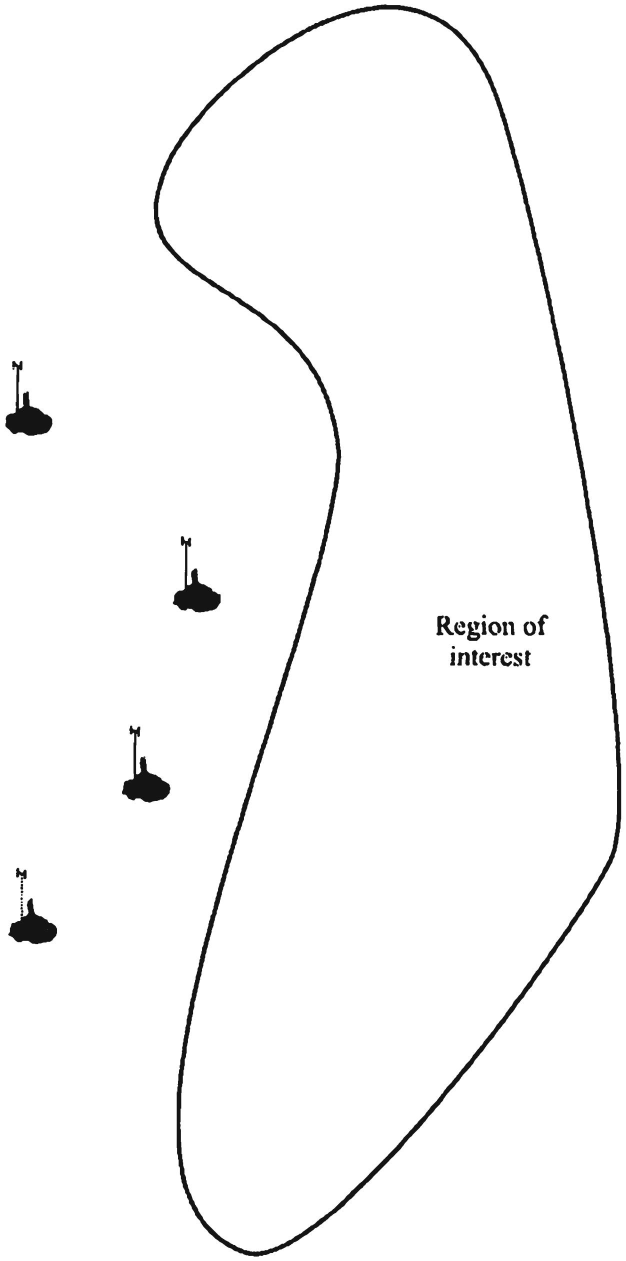


Figure 8.43 If the area of interest is shallow and long, the best configuration for EW systems is a convex baseline as shown here.

overflight of the target area and the geometry shown in Figure 8.45 is possible. Coverage areas are extended due to the overflight and therefore close proximity of the UAS. In addition, since it is closer to the targets, whatever LOP inaccuracy is exhibited is somewhat ameliorated. Of course, the same GDOP concerns as discussed above are still present on the baselines between any two of the systems shown.

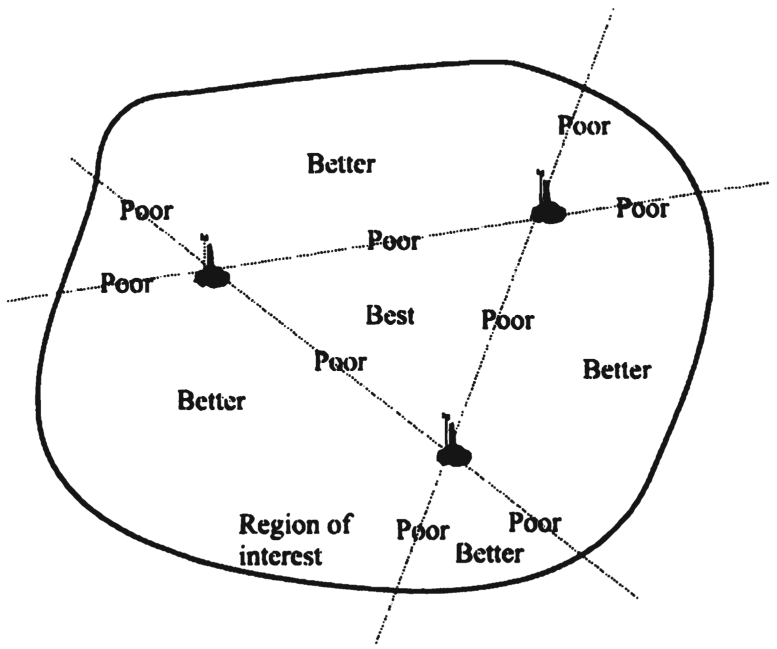


Figure 8.44 Expected quality of direction-finding coverage with three ground systems in a nonlinear battlespace.

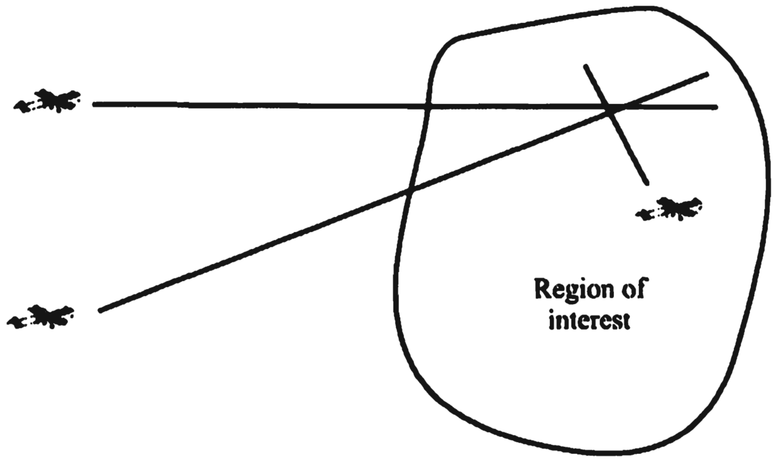


Figure 8.45 Configuration of three airborne direction-finding systems when two are standoff and one is overflight of the region of interest.

8.7 Concluding Remarks

Several techniques for computing an LOP were presented in this chapter. Most of the techniques measure the phase difference at two or more antennas, the time of arrival differences at these antennas, or the amplitude comparisons at two or more antennas. Alternately the Doppler frequency shift of a rotating antenna compared to a stationary antenna is measured. Whatever the technique, the bearing angle to the target is estimated.

Systematic accuracy of practical implementations of these approaches can be as good as 1° or 2° or less when calibration of the antennas is performed. Operational accuracies are typically somewhat worse than this due to the various operational scenarios involved. Three position-fixing algorithms were also presented that utilize the LOPs from two or more collection sites to compute the fix of the target.

In the HF frequency range where long-range communications rely on reflections of the signals from the ionosphere, it is possible to compute a location from a single system as long as the height of the reflection point in the ionosphere is known or estimated. This technique measures the range and azimuth to the target. Accuracy of doing so is about 10% of range. Alternately or in addition, two or more of these systems can be netted just as in the higher-frequency ranges.

References

[1] Lim, J. S., C.-G. Jung, and G. S. Chae, "A Design of Precision RF Direction Finding Device Using Circular Interferometer," *Proceedings of 2004 Intelligent Signal Processing and Communication Systems*, 2004, ISPACS, November 18–19, 2004, pp. 713–716.

[2] Messer, H., G. Singal, and L. Bialy, "On the Achievable DF Accuracy of Two Kinds of Active Interferometers," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 32, No. 3, July 1996, pp. 1158–1164.

[3] Pender, H., and K. McIlwain, *Electrical Engineers Handbook: Electric Communication and Electronics*, 4th Ed., New York: John Wiley & Sons, 1963, p. 1–16.

[4] Torrerri, D. J., *Principles of Secure Communication Systems*, 2nd Ed., Norwood, MA: Artech House, 1992, p. 353.

[5] Struckman, K., "Correlation Interferometer Geolocation," *Proceedings IEEE MILCOM* 2006, pp. 1141–1144.

[6] Balogh, L., and I. Kollar, "Angle of Arrival Estimation Based on Interferometer Principle," *Proceedings WISP* 2003, Budapest, September 4–6, 2003, pp. 219–223.

[7] Xiaobo, A., and F. Zhenghe, "A Single Channel Correlative Interferometer Direction Finder Using VXI Receiver," *Proceedings IEEE 2002 3rd International Conference on Microwave and Millimeter Wave Technology*, pp. 1158–1161.

[8] Park, C.-S, and D.-Y Kim, "The Fast Correlative Interferometer Direction Finder Using I/Q Demodulator, " *Proceedings IEEE Asia-Pacific Conference on Communications*, 2006, APCC 2–6, August 2006, pp. 1–5.

[9] Wiley, R. G., *Electronic Intelligence: The Interception of Radar Signals*, Dedham, MA: Artech House, 1985, p. 102.

[10] Gregoire, D. G., and G. B. Singletary, "Advanced ESM AOA and Location Techniques," *Proceedings of the IEEE 1989 National Aerospace and Electronics Conference*, NAECON 1989, Volume 2, May 22–26 1989, pp. 917–924.

[11] Brinegar, C., "Passive Direction Finding: Combining Amplitude and Phase Based Methods," *Proceedings of the IEEE 2000 National Aerospace and Electronics Conference*, NAECON 2000, October 10–12 2000, pp. 78-84.

[12] Peavy, D., and T. Ogunfunmi, "The Single Channel Interferometer Using a Pseudo-Doppler Direction Finding System," *Proceedings 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing,* ICASSP-97, Volume 5, April 21–24, 1997, pp. 4129–4132.

[13] Gething, P. J. D., *Radio Direction Finding and Superresolution*, London: Peter Peregrinus Ltd., 1991, Ch. 3.

[14] Schmidt, R. O., "Multiple Emitter Location and Signal Parameter Estimation," *Proceedings of the RADC Spectrum Estimation Workshop*, Griffiths Air Force Base, Rome, NY, 1979, pp. 243–258; republished in *IEEE Transactions on Antennas and Propagation*, Vol. AP-34, No. 3, March 1986, pp. 276–280.

[15] Gething, P. J. D., *Radio Direction Finding and Superresolution*, London: Peter Peregrinus Ltd., 1991, p. 217.

[16] Gething, P. J. D., *Radio Direction Finding and Superresolution*, London: Peter Peregrinus Ltd., 1991, Ch. 1.

[17] Warrington, E. M., P. Hamadyk, and T. B. Jones, "Direction of Arrival Measurements of Oblique Chirp Sounder Signals at a High Latitude Site," *Eighth IEEE International Conference on Antennas and Propagation*, 1993, pp. 492–495.

[18] Jones, T. B., and E. M. Warrington, "Direction of Arrival Measurements of HF Signals Propagated over High Latitude Paths," *IEE Colloquium on High Latitude Ionospheric Propagation*, April 29, 1992, pp. 3/1–3/8.

[19] Warrington, E. M., and T. B. Jones, "Observations of the Direction of Arrival of HF Signals Propagated over High Latitude and Mid Latitude Paths," *Eighth IEEE International Conference on Antennas and Propagation*, 1993, pp. 419–422.

[20] Schmidt, R. O., "Least Squares Range Difference Location," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 32, No. 1, Jan 1996, pp. 234–242.

[21] Gething, P. J. D., *Radio Direction Finding and Superresolution*, London: Peter Peregrinus Ltd., 1991, p. 316.

[22] Stansfield, R. G., "Statistical Theory of D. F. Fixing," *Journal, Institute of Electrical Engineers*, Vol. 94, Part IIIa, 1947, pp. 762–770.

[23] Brown, R. M., *Emitter Location Using Bearing Measurements from a Moving Platform*, NRL Report 8483, Naval Research Laboratory, Washington, D.C., June 1981.

[24] Legendre, A., *Nouvelles Methods Pour la Determination des Orbites des Cometes*, Paris, 1805, pp. 72–75.

[25] Sage, A. P., and J. L. Melsa, *Estimation Theory with Applications to Communications and Control*, New York: John Wiley & Sons, 1983, pp. 237–239 and pp. 244–245.

[26] Foy, W. H., "Position Location Solutions by Taylor Series Estimation," *IEEE Transactions on Aerospace and Electronic Systems*, March 1976, pp. 187–193.

[27] Blachman, N. M., "On Combining Target-Location Ellipses," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. AES-27, No. 2, March 1989, pp. 284–287.

[28] Roecker, J. A., "On Combining Multidimensional Target Location Ellipsoids," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 27, No. 1, January 1991, pp. 175–176.

[29] Davis, K., "Ionospheric Radio Propagation," NBS Monograph 80, U.S. Government Printing Office, 1965, pp. 161–162.

[30] Torrerri, D. J., "Statistical Theory of Passive Location Systems," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. AES-20, No. 2, March 1984, pp. 183–197.

Chapter 9

Quadratic Position-Fixing Techniques

9.1 Introduction

When the TOA, the TDOA (denoted here by τ), and/or the frequency difference of arrival ($\dot{\tau}$), or differential Doppler by another appellation, measured at two or more widely dispersed and moving sensors, are used to geolocate emitters, then quadratic LOPs result. Generally, the intersection of these curves is taken as the emitter location. In order for there to be frequency differences caused by movement, either the sensor (one or more) or the target, or both, must be in motion. Therefore, such sensor systems are typically mounted on airborne platforms. For those interested in more details about quadratic PF processing, [1–19] are recommended.

9.2 Time Difference of Arrival

The arrival time of a signal at two or more dispersed sensors can be used to estimate the geographic location of the emitting target [20]. The geometry is shown in Figure 9.1, where, for simplicity, only two receiving systems are shown. Since the transmitter and/or the receiving systems can be elevated, r_1 and r_2 are slant ranges between the transmitter and the sensor systems.

These distances can be expressed as

$$r_i = ct_i \qquad\qquad i = 1,2 \tag{9.1}$$

where c is the speed of propagation of the signal, normally for communication signals in air assumed to be the speed of light, and t_i is the time between when the signal leaves the transmitter and when it arrives at the sensor. The τ is the time

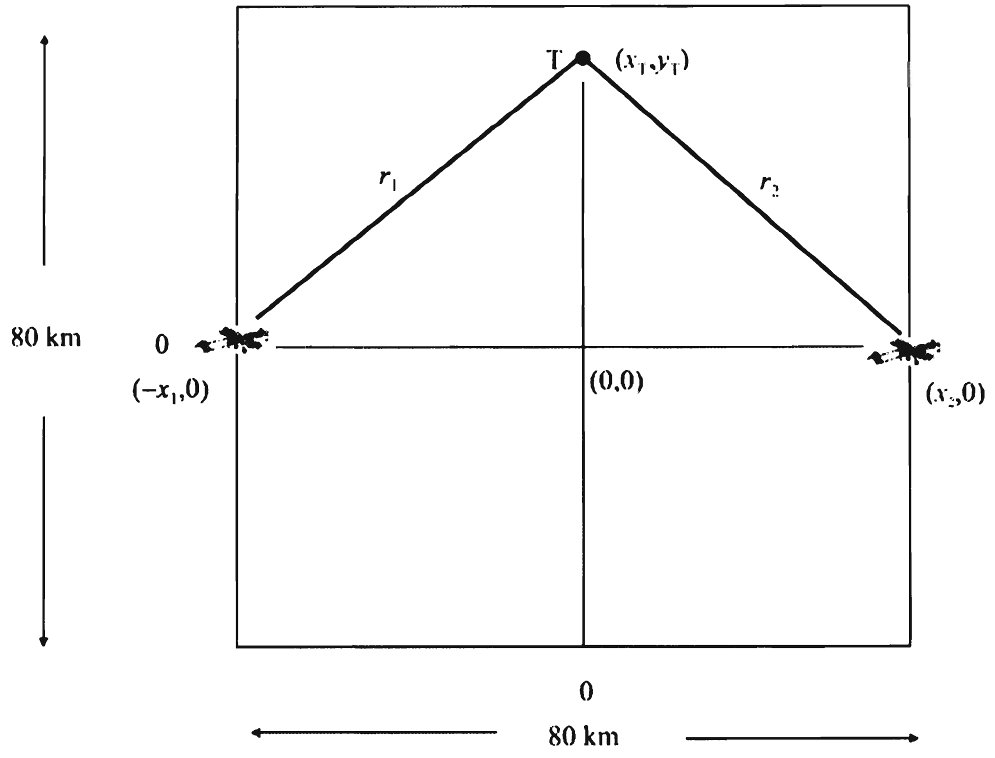


Figure 9.1 Example scenario for illustrating the concepts of τ and $\dot{\tau}$ emitter location.

difference between when the signal arrives at one receiving site and the other, namely,

$$\tau = t_2 - t_1 = \frac{r_2}{c} - \frac{r_1}{c} = \frac{1}{c}(r_2 - r_1) \tag{9.2}$$

From Figure 9.1,

$$r_i = \sqrt{(x_\mathrm{T} - x_i)^2 + y_\mathrm{T}^2} \qquad i = 1,2 \tag{9.3}$$

so that

$$\tau = \frac{1}{c}\left[\sqrt{(x_\mathrm{T} - x_1)^2 + y_\mathrm{T}^2} - \sqrt{(x_\mathrm{T} - x_2)^2 + y_\mathrm{T}^2}\right] \tag{9.4}$$

The curve defined by this expression is a parabola. Since t_2 can be less than or greater than t_1, the τ can be greater than or less than zero. The τ contour, (9.4), for the scenario shown in Figure 9.1 is shown in Figure 9.2. The τ varies between

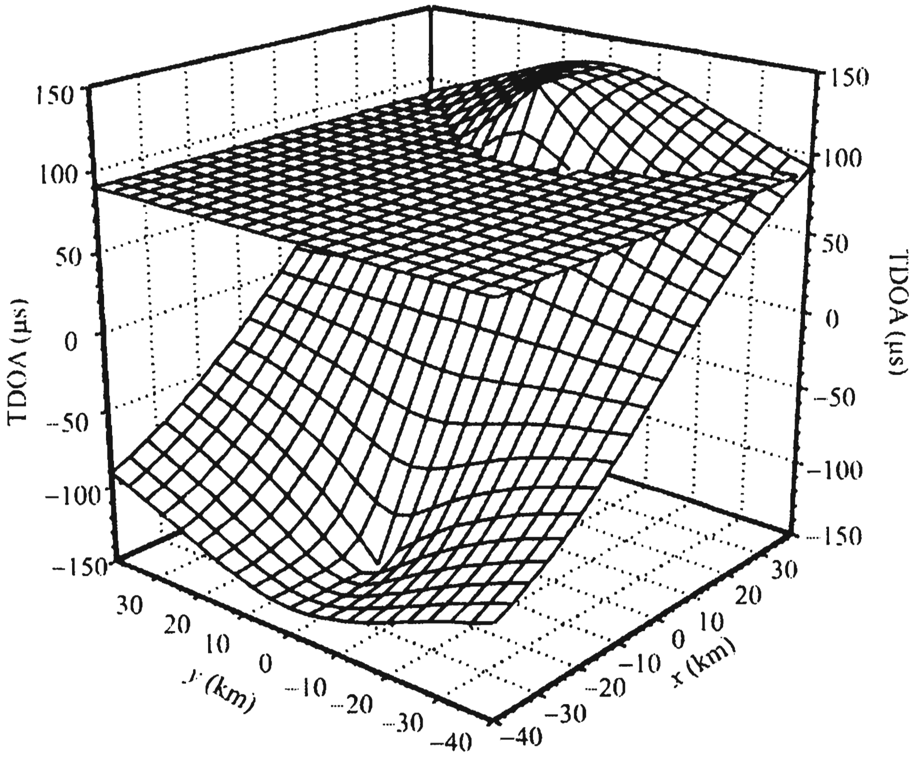


Figure 9.2 Intersection of the τ curve with the plane defined by a constant τ curve of 90 μs forms a parabolic curve.

about −130 μs to +130 μs. Also shown is a constant τ curve which is a plane; in this case τ = 90 μs. When converted to distances, the intersection of these curves forms a hyperbola. As visualized from above, several of these curves are shown in Figure 9.3.

Clearly only two sensor systems cannot produce a unique geolocation instantaneously. To accomplish this at least three sensor systems are required. Suppose there are S sensors available for computation of the position fix [21]. As above, the governing equations are given by

$$d_i - d_j = \|\mathbf{r}_i - \mathbf{r}_\mathrm{T}\| - \|\mathbf{r}_j - \mathbf{r}_\mathrm{T}\| = c(t_i - t_j) = ct_{i,j} \qquad i,j = 0,1,\cdots,S-1 \qquad (9.5)$$

for all pairs of sensors, (i, j) where $\mathbf{r}_i$ corresponds to the vector from the origin to sensor i and $\mathbf{r}_\mathrm{T}$ corresponds to the vector from the origin to the target.

Without loss of generality, assume that all of the arrival times are compared with the arrival time at a sensor located at coordinates (0, 0) as shown in Figure 9.4. Thus, the time differences of arbitrary (i, j) are not used, just $(i, 0)$ for all i. The equation in this case reduces to (because $r_j = 0$)

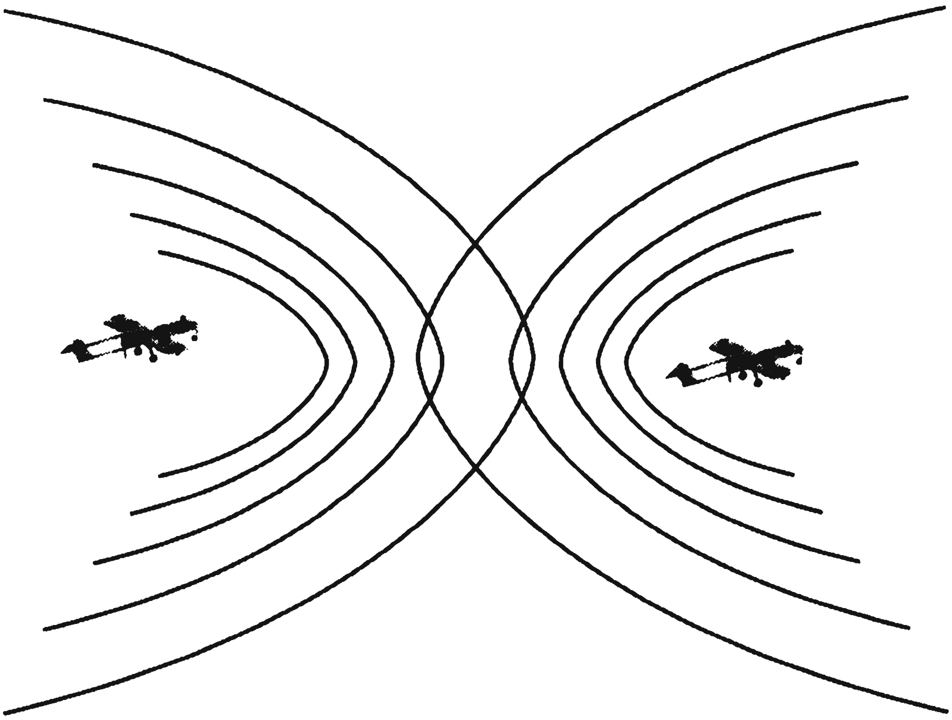


Figure 9.3 Lines of constant *d* based on constant τ are hyperbolas.

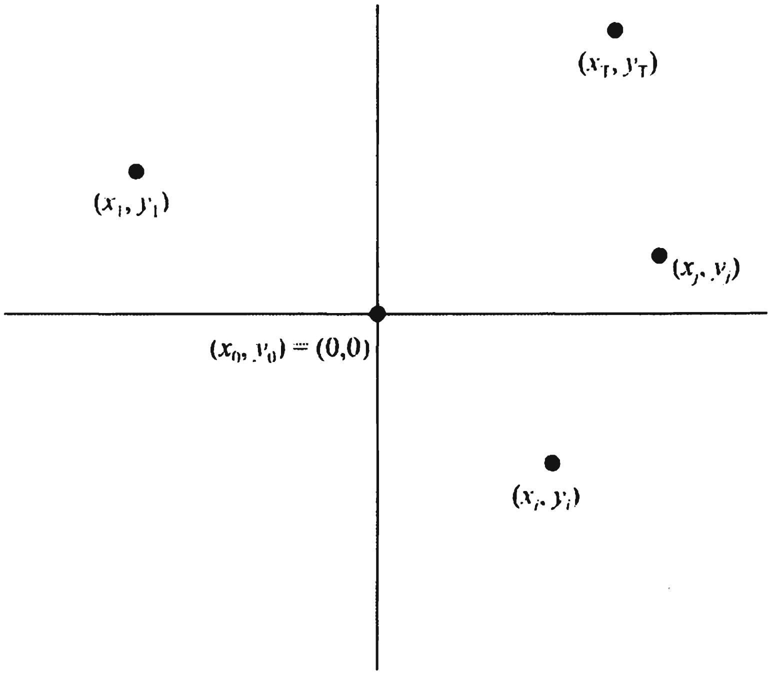


Figure 9.4 Sensor grid and target in two dimensions.

$$d_i = \|\mathbf{r}_i - \mathbf{r}_\mathrm{T}\| = c\left(t_{i,0} + t_0\right) \tag{9.6}$$

Squaring (9.6) yields

$$\left(x_i - x_\mathrm{T}\right)^2 + \left(y_i - y_\mathrm{T}\right)^2 + \left(z_i - z_\mathrm{T}\right)^2 = c^2\left(t_{i,0} + t_0\right)^2 \tag{9.7}$$

Expanding

$$\begin{aligned} x_i^2 - 2x_i x_\mathrm{T} + x_\mathrm{T}^2 + y_i^2 - 2y_i y_\mathrm{T} + y_\mathrm{T}^2 + z_i^2 - 2z_i z_\mathrm{T} + z_\mathrm{T}^2 \\ = c^2\left(t_{i,0}^2 + 2t_{i,0} + t_0^2\right) \end{aligned} \tag{9.8}$$

At the reference site at (0, 0)

$$\|\vec{x}\|^2 = \vec{x}^\mathrm{T}\vec{x} = x_\mathrm{T}^2 + y_\mathrm{T}^2 + z_\mathrm{T}^2 = c^2 t_0^2 \tag{9.9}$$

Subtracting (9.9) from (9.8)

$$x_i^2 + y_i^2 + z_i^2 - 2\left(x_i x_\mathrm{T} + y_i y_\mathrm{T} + z_i z_\mathrm{T}\right) = c^2 t_{i,0}^2 + 2c^2 t_{i,0} \tag{9.10}$$

or

$$\begin{bmatrix} x_i & y_i & z_i \end{bmatrix} \begin{bmatrix} x_\mathrm{T} \\ y_\mathrm{T} \\ z_\mathrm{T} \end{bmatrix} + ct_{i,0}\sqrt{x_\mathrm{T}^2 + y_\mathrm{T}^2 + z_\mathrm{T}^2} = \sum -\frac{1}{2}c^2 t_{i,0}^2 + \frac{1}{2}\left(x_i^2 + y_i^2 + z_i^2\right) \tag{9.11}$$

which utilized the identity

$$ct_0 = \sqrt{x_\mathrm{T}^2 + y_\mathrm{T}^2 + z_\mathrm{T}^2} \tag{9.12}$$

Putting this in matrix form

$$\vec{p}_i \vec{p}_\mathrm{T} + ct_{i,0}\|\vec{x}_\mathrm{T}\| = -\frac{1}{2}ct_{i,0}^2 + \frac{1}{2}\|\vec{p}_i\|^2 \tag{9.13}$$

where $\vec{p}_i$ is the position vector of sensor i such that $\vec{p}_i = \begin{bmatrix} x_i & y_i & z_i \end{bmatrix}^{\mathrm{T}}$ and $\vec{p}_{\mathrm{T}}$ is the position vector of the target, $\vec{p}_{\mathrm{T}} = \begin{bmatrix} x_{\mathrm{T}} & y_{\mathrm{T}} & z_{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$, in three dimensions.

Extending this result for all of the $i = 1,\ldots, S$ sensors yields

$$\mathbf{P}\vec{p}_{\mathrm{T}} + \vec{c}\left\|\vec{p}_{\mathrm{T}}\right\| = \vec{d} \tag{9.14}$$

where

$$\mathbf{P} = \begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ \vdots & \vdots & \vdots \\ x_{S-1} & y_{S-1} & z_{S-1} \end{bmatrix} \quad \vec{p}_{\mathrm{T}} = \begin{bmatrix} x_{\mathrm{T}} \\ y_{\mathrm{T}} \\ z_{\mathrm{T}} \end{bmatrix}$$

$$\vec{c} = c\vec{t} \quad \vec{t} = \begin{bmatrix} t_{1,0} \\ t_{2,0} \\ \vdots \\ t_{S-1,0} \end{bmatrix} \tag{9.15}$$

$$\vec{d} = \frac{1}{2}\operatorname{diag}\left(\mathbf{P}\mathbf{P}^{\mathrm{T}} - c^2\vec{t}\vec{t}^{\mathrm{T}}\right)$$

When $S = 4$ and $\mathbf{P}$ is nonsingular, then $\vec{p}_{\mathrm{T}}$ can be determined as

$$\vec{p}_{\mathrm{T}} = \mathbf{P}^{-1}\left(\vec{d} - \vec{c}\left\|\vec{p}_{\mathrm{T}}\right\|\right) \tag{9.16}$$

Substituting (9.3) into (9.1) yields

$$\left\|\vec{p}_{\mathrm{T}}\right\|^2 = \vec{p}^{\mathrm{T}}\vec{p} = \left\{\mathbf{P}^{-1}\left(\vec{d} - \vec{c}\left\|\vec{p}_{\mathrm{T}}\right\|\right)\right\}^{\mathrm{T}}\left\{\mathbf{P}^{-1}\left(\vec{d} - \vec{c}\left\|\vec{p}_{\mathrm{T}}\right\|\right)\right\} \tag{9.17}$$

Let $a = \left\|\vec{p}_{\mathrm{T}}\right\|$ and $\mathbf{Q} = (\mathbf{P}\mathbf{P}^{\mathrm{T}})^{-1}$; then this expression reduces to

$$\left(\vec{c}^{\mathrm{T}}\mathbf{Q}\vec{c} - 1\right)a^2 - 2\vec{d}^{\mathrm{T}}\mathbf{Q}\vec{c}a + \vec{d}^{\mathrm{T}}\mathbf{Q}\vec{d} = 0 \tag{9.18}$$

which is a quadratic equation that can easily be solved for *a* which is the range of the target from the origin. Substituting this range back into (9.14) will solve the problem for the target location. Two results ensue, so other information must be used to determine which is correct. Frequently one answer is negative and since *a* represents the range to the target, the positive root is the correct one. Otherwise, other information must be employed.

In general $S > 4$, so this is an over-determined system of equations. Instead of the inverse, then, the pseudo-inverse

$$\mathbf{P}^{\diamond} = (\mathbf{P}^{\mathrm{T}}\mathbf{P})^{-1}\mathbf{P}^{\mathrm{T}} \tag{9.19}$$

is used. The pseudo-inverse solves (9.14) in the least squares sense (see Section 11.2.6.1).

The Cramer-Rao bound on parameter estimation is a frequently used measure of how well such a parameter can be measured. Under some reasonable assumptions it represents the best that can be obtained under those assumptions. The Cramer-Rao bound for estimating the TOA of a signal at a sensor is given by [22]

$$\sigma_{\iota}^{2} \geq \left(\frac{2E}{N_0}\beta^2\right)^{-1} \tag{9.20}$$

where σ_{ι}^{2} is the variance of the measurement and is given in seconds squared; in this case, E is the energy in the signal (watt-seconds, for example), N_0 is the level of noise per unit bandwidth present (watts per hertz, for example), and β is a measure of the bandwidth occupied by the signal, measured in radians. This equation reduces to

$$\sigma_{\iota} = \frac{1}{\beta}\frac{1}{\sqrt{BT\gamma}} \tag{9.21}$$

where B is the noise bandwidth of the receivers, T is the integration time, and γ is the effective input SNR at the two sensor sites.

β is the rms radian frequency, which is a measure of the bandwidth of the signal and is given by

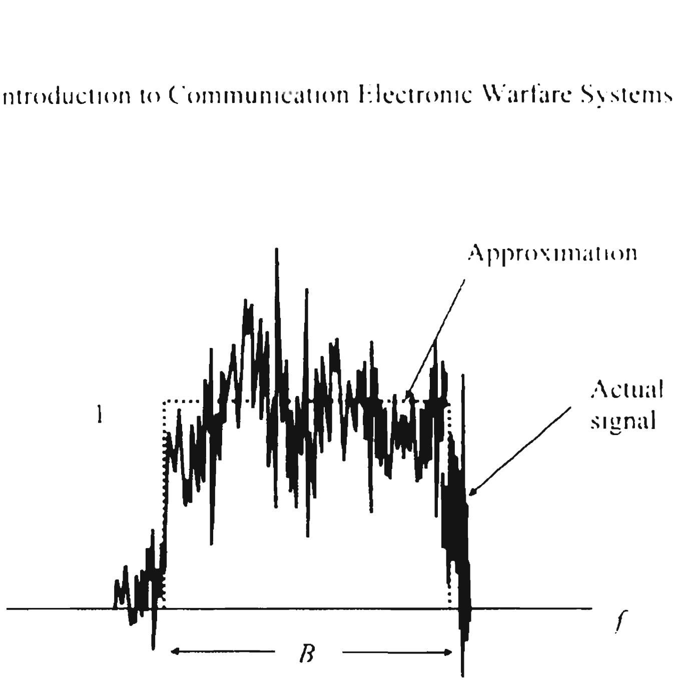


Figure 9.5 Example of an ideal signal that approximates a communication signal.

$$\beta = 2\pi \left[\frac{\int_{-\infty}^{\infty} f^2 |S(f)|^2 \, df}{\int_{-\infty}^{\infty} |S(f)|^2 \, df} \right]^{1/2} \tag{9.22}$$

where $S(f)$ is the spectrum of the signal. For an ideal signal with sharp edges in the frequency spectrum as shown in Figure 9.5, $\beta = \pi B/\sqrt{3}$ (where B is in hertz), for example.

Thus, γ is a composite SNR at the two sensors. If γ_1 and γ_2 are the SNRs at the two sensors, then γ is given by

$$\frac{1}{\gamma} = \frac{1}{2}\left[\frac{1}{\gamma_1} + \frac{1}{\gamma_2} + \frac{1}{\gamma_1\gamma_2}\right] \tag{9.23}$$

Note that there has been no assumption made about how the TDOA is determined. It has been shown that phase data can be used to estimate the TDOA [23, 24]. In some cases the signal contains a feature that occurs at specific times and at which the time can be measured. This time of arrival (TOA) can be compared to a reference sensor and the TDOA between the sensors determined. Pulsed radar, for example, has a leading edge to the radar pulse that can be used. Other signals may have specific data patterns inserted to be used for time measurements. Time delay can even be computed in multipath channels [25–27]. The biggest advantage of such feature-based TOA measurements is the limited amount if information that needs to be exchanged with other sensors in order to compute the pair-wise TDOAs.

An alternate method that often yields accurate time measurements when such features as these are not available is with cross-correlation [27–35]. The cross-correlation of signals from two (at a time) sensors is computed. The TDOA of the two signals is indicated by a peak in the result. The multiple TDOAs that emerge from incorporating the sensors two at a time can then be used to compute the location of the target.

One of the biggest shortcomings of cross-correlation computation for finding the TDOAs is the data bandwidth requirements if the sensors are physically located a distance apart. While it is not always true, crosscorrelation generally requires correlation of two pre-detected signals. They can be converted to digital form, but still the data bandwidth requirements can be quite substantial.

9.2.1 Position-Fixing Using TDOA Measurements

In this section we present an exemplar of a geoposition computation algorithm initially put forth by Bard and Ham [36] and Bard, Ham, and Jones [37]. It is but one of many documented in the literature, but is reasonably concise and straightforward to understand.

The TOA measurements lead to pair-wise TDOA calculations through

$$\Delta t_i = t_i - t_0 \tag{9.24}$$

where t_0 is an unknown quantity representing the TOA of the target signal at the reference sensor site, assumed to be located at $\mathbf{x}_0 = (x_0, y_0, z_0) = (0, 0, 0)$ without loss of generality, and t_i is the arrival time of the target signal at sensor site i. In three-dimensions the arrival of the signal at the ith site is governed by

$$(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2 = c^2(\Delta t_i + t_0)^2 \tag{9.25}$$

where c is the speed of light, $\mathbf{x} = (x, y, z)$ is the position of the target, and $\mathbf{x}_i = (x_i, y_i, z_i)$ is the location of the ith sensor. At the reference site

$$x^2 + y^2 + z^2 = c^2 t_0^2 = r_0^2 = \|\vec{x}\|^2 = \vec{x}^{\mathrm{T}}\vec{x} \tag{9.26}$$

Subtract (9.26) from (9.25) to yield

$$\begin{aligned}&\begin{bmatrix}x_i & y_i & z_i\end{bmatrix}\begin{bmatrix}x\\ y\\ z\end{bmatrix}+c\Delta t_i\sqrt{x^2+y^2+z^2}\\ &=-\frac{1}{2}c^2\Delta t_i^2+\frac{1}{2}\left(x_i^2+y_i^2+z_i^2\right), \qquad i=1,2,\ldots,M\end{aligned} \tag{9.27}$$

where M is the number of remote sensors (including the reference site, there are $M+1$ sensors). In vector form (9.27) becomes

$$\vec{x}_i\vec{x}+c\Delta t_i\left\|\vec{x}\right\|=-\frac{1}{2}c^2\Delta t_i^2+\frac{1}{2}\left\|\vec{x}_i\right\|^2 \tag{9.28}$$

We compute the gradient of the measurements with respect to the unknown position coordinates as

$$\frac{\partial}{\partial\vec{x}}\vec{x}_i\vec{x}+c\frac{\partial}{\partial\vec{x}}\Delta t_i\left\|\vec{x}\right\|=-\frac{1}{2}c^2\frac{\partial}{\partial\vec{x}}\Delta t_i^2 \tag{9.29}$$

Rearranging

$$\vec{x}_i^{\mathrm{T}}+c\left(\Delta t_i\frac{\vec{x}}{\left\|\vec{x}\right\|}+\left\|\vec{x}\right\|\frac{\partial\Delta t_i}{\partial\vec{x}}\right)=-c^2\Delta t_i\frac{\partial\Delta t_i}{\partial\vec{x}} \tag{9.30}$$

The solution of (9.30) is

$$\frac{\partial\Delta t_i}{\partial\vec{x}}=\frac{1}{c}\left(\frac{\vec{x}-\vec{x}_i^{\mathrm{T}}}{\left\|\vec{x}-\vec{x}_i^{\mathrm{T}}\right\|}-\frac{\vec{x}}{\left\|\vec{x}\right\|}\right) \tag{9.31}$$

Define a gradient matrix as

$$\mathbf{H}=\begin{bmatrix}\left(\dfrac{\partial\Delta t_1}{\partial\vec{x}}\right)^{\mathrm{T}} & \left(\dfrac{\partial\Delta t_2}{\partial\vec{x}}\right)^{\mathrm{T}} & \cdots & \left(\dfrac{\partial\Delta t_M}{\partial\vec{x}}\right)^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}} \tag{9.32}$$

Equation (9.27) defines the ith row of the matrix equation as

$$\mathbf{X}\vec{x} + c\left\|\vec{x}\right\|_2 = \vec{m} \tag{9.33}$$

where

$\mathbf{X} = \begin{bmatrix} \vec{x}_1^{\mathrm{T}} & \vec{x}_2^{\mathrm{T}} & \cdots & \vec{x}_M^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} \in \Re^{M\times 3}$ $\Delta\vec{t}$ = time difference measurement vector

$\vec{c} = c\Delta\vec{t} \in \Re^{M\times 1}$ $\vec{m} = \left[\frac{1}{2}\left(\mathbf{X}\mathbf{X}^{\mathrm{T}} - \vec{c}\vec{c}^{\mathrm{T}}\right)\right]_{ii}$, $i = 1, 2, \ldots, M \in \Re^{M\times 1}$

$\left\|\bullet\right\|_2$ is the L_2 norm. Equation (9.33) is the TDOA equation. Formulate

$$\left(\vec{c}^{\mathrm{T}}\mathbf{\Phi}\vec{c} - 1\right)r_o^2 - 2\vec{m}^{\mathrm{T}}\mathbf{\Phi}cr_0 + \vec{m}^{\mathrm{T}}\mathbf{\Phi}\vec{m} = 0 \tag{9.34}$$

where

$$\mathbf{\Phi} = \mathbf{Q}^{-1}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Q}^{-1} \tag{9.35}$$

and **Q** is the covariance matrix of the noise:

$$\mathbf{Q} = \mathcal{E}\{\vec{\eta}\vec{\eta}^{\mathrm{T}}\} \tag{9.36}$$

with

$$\mathcal{E}\{\vec{\eta}\} = 0 \tag{9.37}$$

The solution of (9.34) yields r_0 which is used to compute the target location with

$$\hat{\vec{x}}(r_0) = \left(\mathbf{X}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Q}^{-1}\left(\vec{m} - \vec{c}r_0\right) \tag{9.38}$$

The target location estimate thus obtained can be refined by using a Taylor series expansion around $\hat{\vec{x}}$ as

$$\Delta t_i = f(\vec{x}_i, \vec{x}) = f(\vec{x}_i, \hat{\vec{x}} + \Delta\vec{x}) \tag{9.39}$$

The expansion about $\hat{\vec{x}}$ is

$$\Delta t_i = f(\vec{x}_i,\vec{x}) = f(\vec{x}_i,\hat{\vec{x}}) + \frac{\partial f(\vec{x}_i,\hat{\vec{x}})^{\mathrm{T}}}{\partial \hat{\vec{x}}}\Delta\vec{x} + \cdots \tag{9.40}$$

The derivative in (9.40) is given as the *i*th row of (9.32). We thus get the linear equation (ignoring second-order and higher terms in the Taylor series)

$$c\Delta\vec{t} - f(\vec{x},\hat{\vec{x}}) = \mathbf{H}\Delta\vec{x} \tag{9.41}$$

where

$$f(\vec{x},\hat{\vec{x}}) = \begin{bmatrix} \left\|\hat{\vec{x}} - \vec{x}_1^{\mathrm{T}}\right\| - \left\|\vec{x}\right\| \\ \left\|\hat{\vec{x}} - \vec{x}_2^{\mathrm{T}}\right\| - \left\|\hat{\vec{x}}\right\| \\ \vdots \\ \left\|\hat{\vec{x}} - \vec{x}_M^{\mathrm{T}}\right\| - \left\|\hat{\vec{x}}\right\| \end{bmatrix} \tag{9.42}$$

9.2.2 Geometric Dilution of Precision

Just as for position-fixing using bearing measurements discussed in Chapter 8, there is GDOP associated with TDOA position-fixing [36, 38]. If L denotes the length of a linear array consisting of three sensors with coordinates given by (0, $-L/2$), (0, 0), and (0, $L/2$) then the lower bound of the standard deviation of the TDOA error is given by

$$\sigma_{tL}^2 = \frac{N_0 L^n \exp(\alpha L)}{2\beta^2 T K_E} \tag{9.43}$$

where $N_0/2$ is the 2-sided noise spectral density, n is the propagation exponent discussed in Chapter 4 ($n = 4$ close to the ground), T is the total measurement time, K_{E} is a constant associated with the average signal power at the receiver, R_{S}, defined by

$$R_{\mathrm{S}} = K_{\mathrm{E}} \exp\left(-\alpha \frac{D}{D^n}\right) \tag{9.44}$$

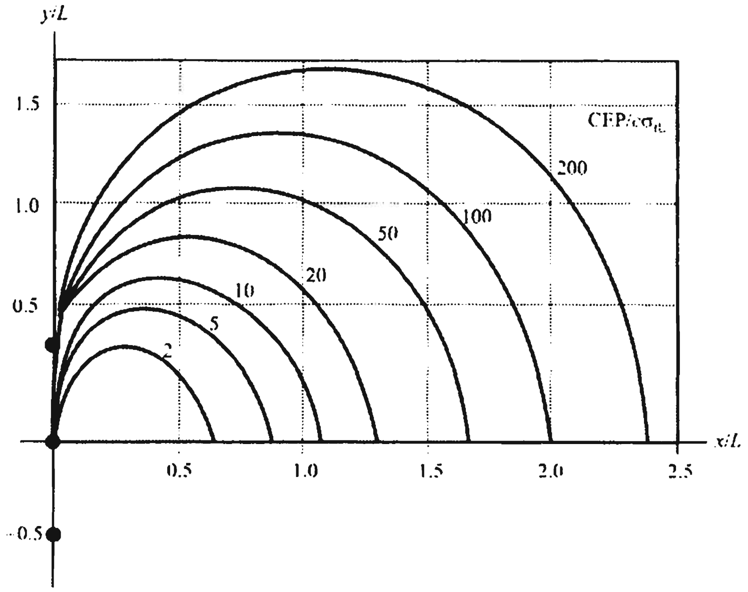


Figure 9.6 TDOA GDOP lines of constant CEP. *c* is the speed of light. (*Source:* [38]. © IEEE 1984. Reprinted with permission.)

and α is a propagation constant. For most frequency ranges of concern to us, $\alpha = 0$. The GDOP for the three-sensor configuration is illustrated in Figure 9.6 when $n = 4$. Care should be exercised when evaluating the error curves when significantly off the perpendicular bisector since the eigenvalues of the measurement correlation matrix can produce numerically unstable computations of the CEP in those cases. As such the CEP is not a very good measure of the accuracy and EEP computations should be used instead.

With the approximation for a communication signal spectrum given in Figure 9.5, using (9.21) as an approximation for σ_{tL}, then at $x/L = 1$ on the perpendicular bisector of the baseline we get

$$\frac{\text{CEP}}{c\sigma_t} \approx 9$$

So with $T = 10$ s, $B = 25$ kHz, and $\gamma = 10$ dB,

$$\text{CEP} \approx 37.7\text{m}$$

demonstrating the accurate position-fixing capability associated with TDOA geolocation technology.

Also in common with bearing position fixing, TDOA geolocation calculations yield unacceptably large errors off the ends of baselines, especially for only three ES systems. This can be seen in Figure 9.6 where the error goes to infinity off each end of the linear baseline.

The *i*th row of the gradient matrix given by (9.32) is equal to the difference between the unit line of the site vector to the *i*th sensor vector and the unit vector to the target. GDOP is defined as

$$\text{GDOP} = \sqrt{\text{trace}\left(\mathbf{H}^{\text{T}}\mathbf{H}\right)^{-1}} \tag{9.45}$$

GDOP is a multiplying factor of the standard deviation of the error. If σ_d denotes the measurement standard deviation then the spherical uncertainty in the calculated geoposition is given by

$$\text{Uncertainty} = \text{GDOP} \times \sigma_d \tag{9.46}$$

For example, if σ_d = 10m (33 ns) on each channel and the GDOP is 30 then the rms spherical uncertainty in that region is 300m.

For a circularly symmetric near-planar array, the GDOP is illustrated in Figure 9.7. When the elevation angle is 2°, *n* in this figure is the number of sensors and varies between 4 and 10.

9.2.3 Time Delay Estimate Limitations

There are limitations to how accurately the time delay can be estimated [39–45]. It has been shown that there is a sharp threshold phenomenology exhibited by time

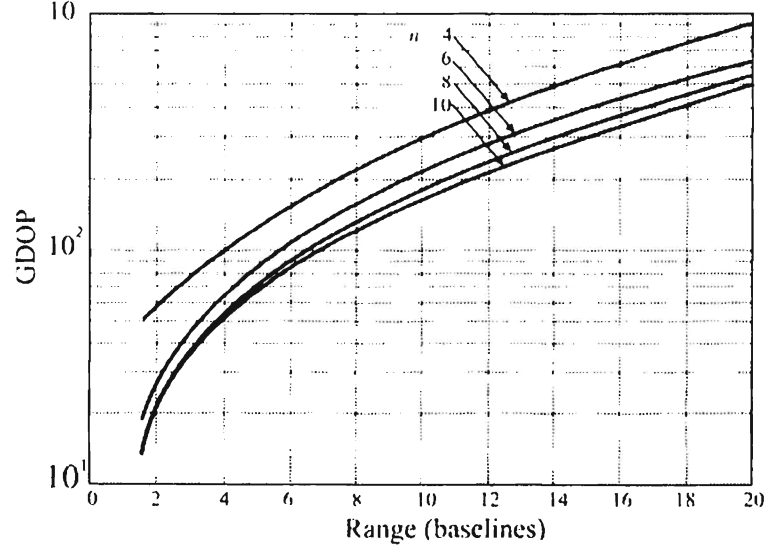


Figure 9.7 TDOA GDOP (Elevation angle = 2°). (*Source:* [36]. © IEEE 1999. Reprinted with permission.)

delay estimators [39, 46]. When the *SNR* is above the threshold the mean-square-error (MSE) is bounded by the Cramer-Rao lower bound (CRLB) given by

$$\mathrm{MSE} \geq \left\{ \frac{T}{2\pi} \int_{\omega_0 - B/2}^{\omega_0 + B/2} 2\omega^2 \mathrm{SNR}(\omega) d\omega \right\}^{-1} = \mathrm{CRLB}\ \text{(above threshold)} \qquad (9.47)$$

where

$$\mathrm{SNR}(\omega) = \frac{\dfrac{S(\omega)}{N_1(\omega)} \dfrac{S(\omega)}{N_2(\omega)}}{1 + \dfrac{S(\omega)}{N_1(\omega)} + \dfrac{S(\omega)}{N_2(\omega)}}$$

For

$$|\omega \pm \omega_0| \leq \frac{B}{2}$$

$S(\omega)$ is the signal received at the two sensors (one a delayed version of the other), and $N_1(\omega)$ and $N_2(\omega)$ are the noise density levels at the two sensors. B is the bandwidth of the signal. When the SNR is below the threshold then the MSE is lower bounded by

$$\mathrm{MSE} \geq \left\{ \frac{T}{2\pi} \int_{\omega_0 - B/2}^{\omega_0 + B/2} 2(\omega - \omega_0)^2 \mathrm{SNR}(\omega) d\omega \right\}^{-1} \quad \text{(below threshold)} \qquad (9.48)$$

For constant SNR, (9.47) becomes

$$\mathrm{MSE} \geq \frac{1}{\left(2\omega_0^2 + \dfrac{B^2}{6} \right) \left(\dfrac{BT}{2\pi} \right) \mathrm{SNR}} \qquad \text{(above threshold)} \qquad (9.49)$$

and (9.48) becomes

$$\text{MSE} \geq \frac{1}{\left(\frac{B^2}{6}\right)\left(\frac{BT}{2\pi}\right)\text{SNR}} \quad \text{(below threshold)} \tag{9.50}$$

The SNR threshold in these expressions is given by

$$\left(\frac{BT}{2\pi}\right)\text{SNR} = 0.28Q^2 \tag{9.51}$$

with

$$Q = \frac{\omega_0}{B} \tag{9.52}$$

The left side of (9.51) is the post-integration SNR. For reasonable values of Q, say above 100, the required post-integration SNR is quite large so therefore for most problems, time delay estimation is performed below threshold.

9.3 Differential Doppler

The frequency difference of arrival, or differential Doppler, for all practical cases of interest herein, where all sensor and/or target velocities are much smaller than the speed of light, is given by

$$\dot{\tau} = \frac{v_2}{c} f_0 - \frac{v_1}{c} f_0 = \frac{f_0}{c}(v_2 - v_1) \tag{9.53}$$

where v_i represents the instantaneous velocity of the sensors relative to the transmitter in the radial direction and f_0 is the frequency of the transmitted signal. Thus,

$$\dot{\tau} = \frac{f_0}{c}\left(\frac{dr_2}{dt} - \frac{dr_1}{dt}\right) \tag{9.54}$$

where r_i is the range between the transmitter and the sensor, as shown in Figure 9.1; thus, dr_i/dt is the rate of change of the range in the radial direction. From (9.3),

$$\frac{dr_i}{dt} = \frac{d\left[\left(x_\mathrm{T} - x_i\right)^2 + y_\mathrm{T}^2\right]^{1/2}}{dt} \tag{9.55}$$

$$= \frac{\left(x_\mathrm{T} - x_i\right)}{\left[\left(x_\mathrm{T} - x_i\right)^2 + y_\mathrm{T}^2\right]^{1/2}} \frac{dx_i}{dt}, \; i = 1, 2 \tag{9.56}$$

where it is assumed that the aircraft are flying at the same speed parallel to the x-axis so that $dy_i/dt = 0$. Denoting $v = v_1 = dx_1/dt = v_2 = dx_2/dt$, then

$$\dot{\tau} = \frac{f_0 v}{c} \left\{ \frac{\left(x_\mathrm{T} - x_1\right)}{\left[\left(x_\mathrm{T} - x_1\right)^2 + y_\mathrm{T}^2\right]^{1/2}} - \frac{\left(x_\mathrm{T} - x_2\right)}{\left[\left(x_\mathrm{T} - x_2\right)^2 + y_\mathrm{T}^2\right]^{1/2}} \right\} \tag{9.57}$$

The surface formed by this expression for the example here is shown in Figure 9.8. Note that the largest differential Doppler frequency is approximately 6.5 Hz or so. This is not much of a frequency difference to try to measure out of 100 MHz. The example considered here uses parameters that are typical for slow-flying applications. That is part of the reason that the differential Doppler values are so low—the velocity of the EW system is small (recall that for zero velocity the differential Doppler is also zero). So the lower the velocity, the lower the differential Doppler component. Faster-moving EW systems improve the measurable values of differential Doppler. The differential Doppler curves are complex quadratic functions. Their form can be seen by examining Figure 9.9 where the above surface is viewed from almost directly overhead and the intersection of the $\dot{\tau} = 4.5$ Hz with the surface contour is shown. When visualized from above, the $\dot{\tau}$ curves look like those in Figure 9.10. Again, the $\dot{\tau}$ from only two sensor platforms do not yield unique fixes. The Cramer-Rao bound for estimating $\dot{\tau}$ is given by

$$\sigma_\mathrm{f} = \frac{1}{T_\mathrm{e}} \frac{1}{\sqrt{BT\gamma}} \tag{9.58}$$

where B, T, and γ are as above, σ_f is the standard deviation of the $\dot{\tau}$ measurement, and T_e is the RMS integration time given by

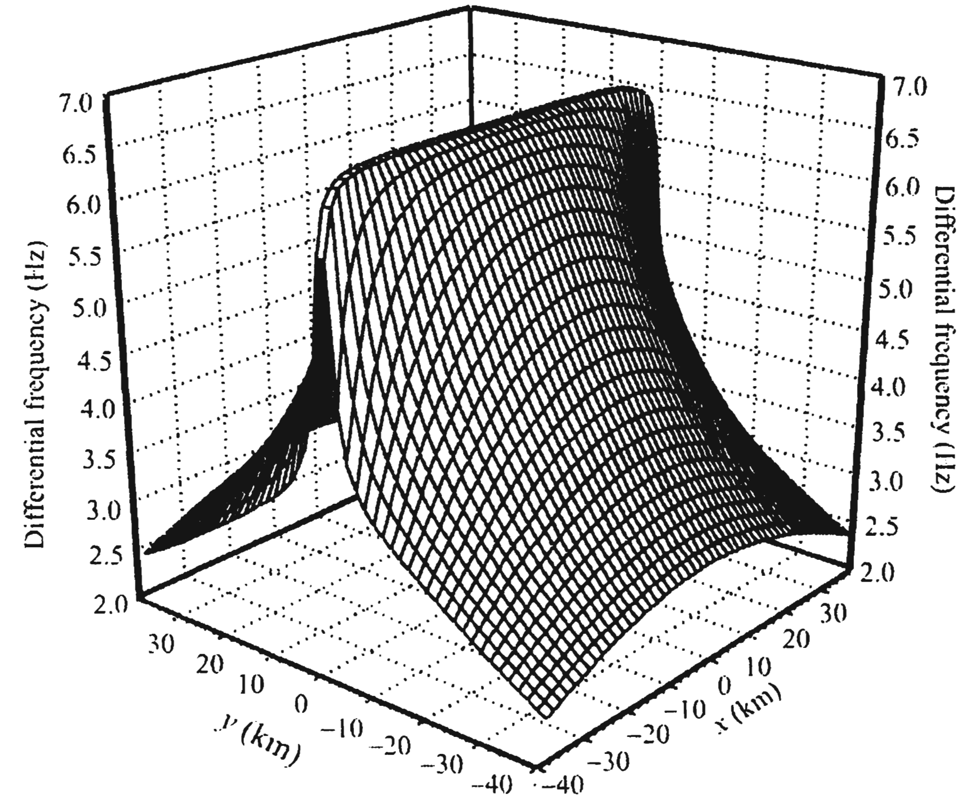


Figure 9.8 Surface of the differential Doppler for the example. In this case v = 10 m/s and f = 100 MHz (λ = 3m).

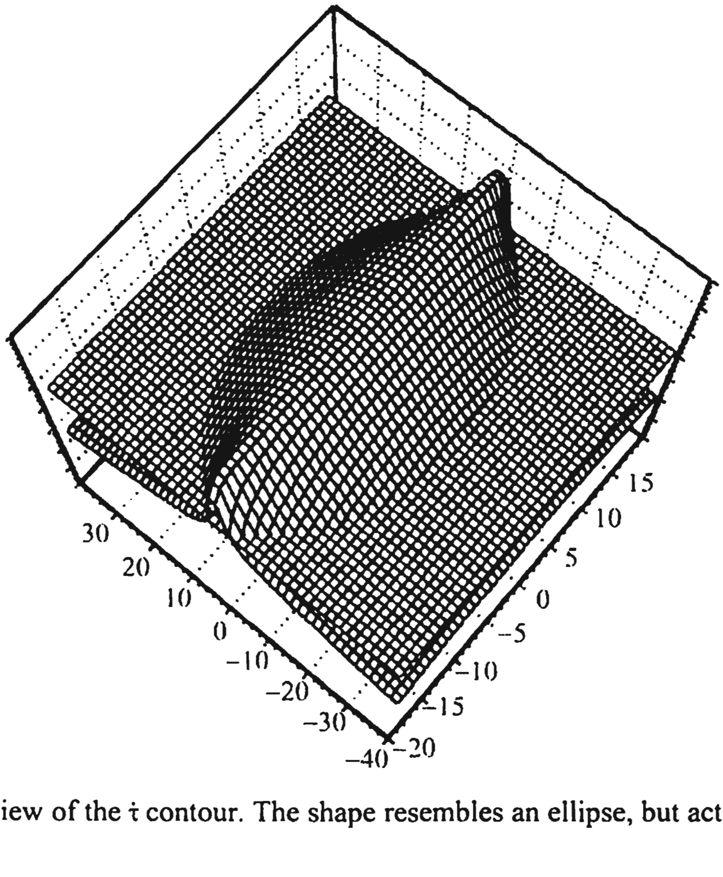


Figure 9.9 Top view of the $\dot{\tau}$ contour. The shape resembles an ellipse, but actually it is a complex quadratic.

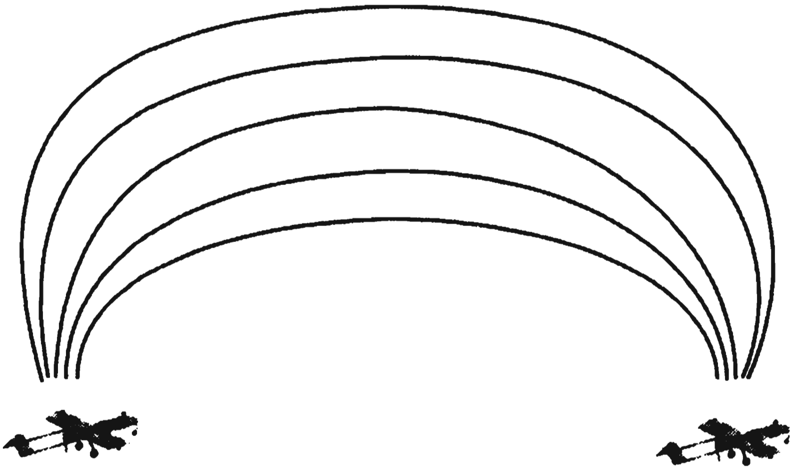


Figure 9.10 Lines of constant $\dot{\tau}$ are complex quadratic curves.

$$T_e = 2\pi \left[\frac{\int_{-\infty}^{\infty} t^2 |u(t)|^2 \, dt}{\int_{-\infty}^{\infty} |u(t)|^2 \, dt} \right]^{1/2} \tag{9.59}$$

where $u(t)$ is the probability density function of the integration time. Again, for example, if the actual integration time is T, then $T_e = \pi T/\sqrt{3}$.

9.4 Cross-Ambiguity Function Processing

An alternative way to compute the τ and $\dot{\tau}$ for communication targets is with the *cross-ambiguity function* (CAF) [4, 11, 47, 48]. This computation simultaneously yields the τ and $\dot{\tau}$ for two sensors, and the CAF must be computed for each pair in order to yield sufficient information to calculate a PF. The CAF is a generalization of the cross-correlation function and is given by

$$\text{CAF}(\tau, t) = \int_0^T s_1(t) s_2(t + \tau) e^{-j\omega t} \, dt \tag{9.60}$$

A three-dimensional graphical plot of the magnitude of the ambiguity function shows the amplitude of the spectrum on one axis, as the dependent variable, versus

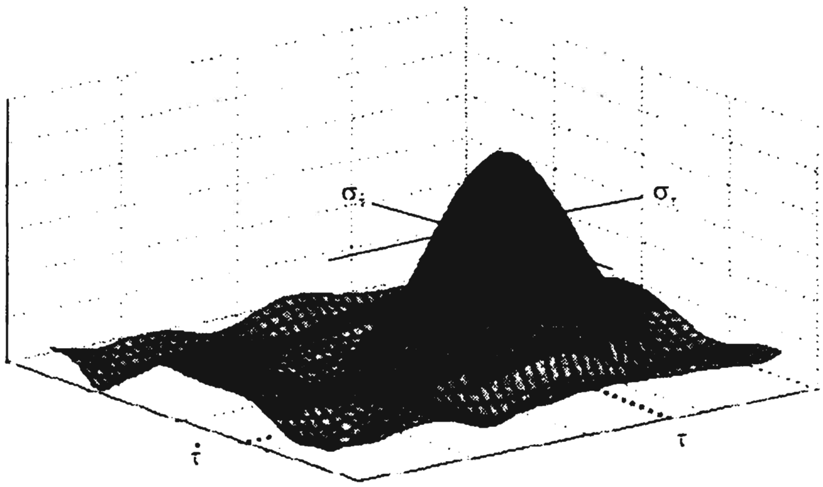


Figure 9.11 Example of the magnitude of a CAF.

the two independent variables on the other two axes, given by the τ and $\dot{\tau}$. With calculation of the ambiguity function, highly precise geolocation of targets can be obtained. A sketch of an example ambiguity function is shown in Figure 9.11. The magnitude of the ambiguity function is searched to ascertain where the highest peak lies. The resultant values of τ and $\dot{\tau}$ where this peak occurs is assumed to be those corresponding to the emitter. The width of the largest peak are measures of the standard deviations of the τ and $\dot{\tau}$ as illustrated in Figure 9.11.

Neither the τ nor the $\dot{\tau}$ yields unique solutions with two sensors. There are an infinite number of points that lead to the same τ and $\dot{\tau}$ values. This is illustrated in Figures 9.3 and 9.11. When the τ and $\dot{\tau}$ are combined, there no longer is an infinite number of points, but there are still more than one. An example solution of computing an emitter location using $\dot{\tau}$ and τ combined is shown in Figure 9.12. With just the two platforms shown, the emitter could be at either point 1 or 2 and the computed solutions would be the same.

There are ways to resolve this ambiguity, one of which is to add a third sensor. Alternately, as long as the emitter stays transmitting long enough, one or both of the two sensors could move to another location and the process repeated, yielding a unique solution as illustrated in Figure 9.13. The contours computed with the second baseline would cross at either point 1 or 2, but not both. The second sensor moved to another position, or there is a third sensor in the formation, so that a second baseline can be formed. In this example, the emitter is at location 1, and that is where the contours intersect for both baselines, whereas for position 2, the intersection of the contours is different. The tradeoffs here are time or a third sensor platform.

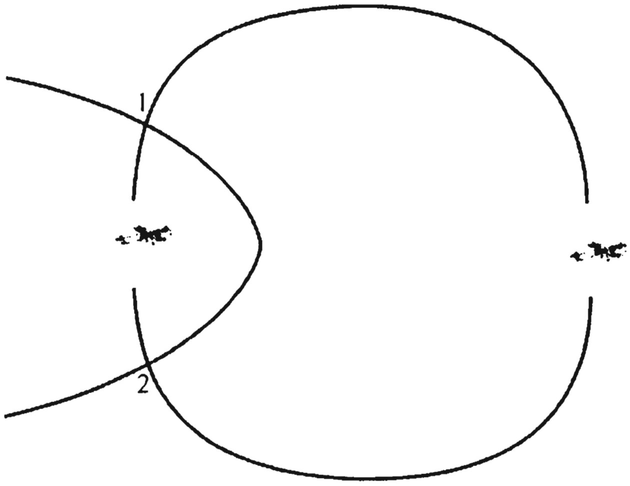


Figure 9.12 Combining the τ solution with the $\dot{\tau}$ solution does not yield a unique location with two platforms.

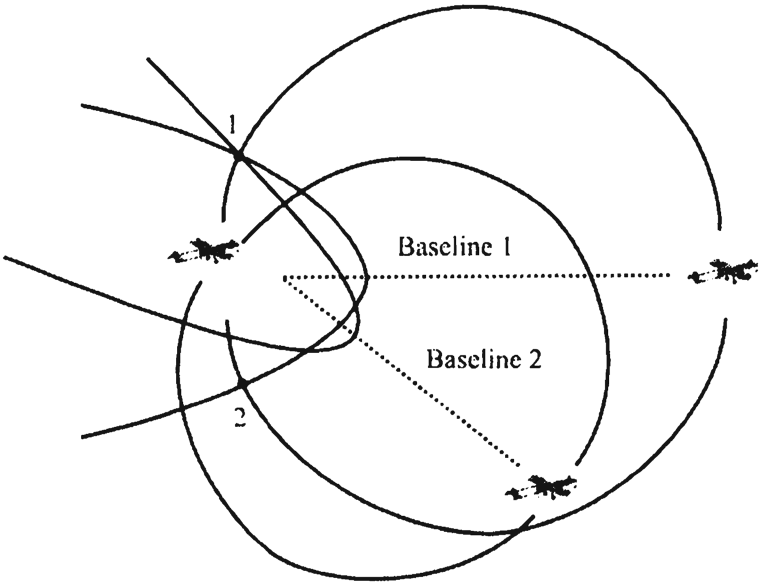


Figure 9.13 Adding a second baseline removes the ambiguity inherent with only a single baseline.

Of course, adding a third system to the configuration adds more than one more baseline. Only two baselines are needed, however, for the process just described. Operationally, the best baselines to use would be determined by the configuration relative to the location of the target. Some pairs of baselines would produce better results than others.

Another way to resolve the ambiguities is to use other DF techniques to indicate which computed location corresponds to the target. Such techniques are typically less accurate than $\tau / \dot{\tau}$ and would therefore not normally be used in the actual calculation of the BPE.

9.4.1 Position-Fix Accuracy

As an example, suppose B = 25 kHz, the channel width in the low VHF frequency band. Further suppose that γ_1 = 10 dB (10) and γ_2 = 15 dB (32). Let T = 100 ms Then $\beta = 90.7 \times 10^3$ radians per seconds, T_e = 0.363 second, and γ = 9. Therefore, σ_t = 63 ns and σ_f = 16.1 mHz. To put this in perspective, signals propagate at about the speed of light, 3×10^8 m/s. Thus σ_t of this magnitude equates to an area of uncertainty in the τ dimension of about 20m. In the Doppler direction the interpretation is not quite so straightforward.

While subject to the same detrimental factors as interferometers and other DF systems, especially the degradation due to multipath reflections, particularly in the ground applications, τ and $\dot{\tau}$ systems have been demonstrated to achieve considerably better accuracy than typical DF systems. This is accomplished, however, at the expense of the time necessary to obtain results. The integration time is a factor in the denominator in both of the above equations. Thus, increasing this time will decrease the standard deviation of the resulting position-fix. It is not unusual for the integration time to increase by two orders of magnitude to obtain such precision results.

The standard deviation of the measurements of the fix coordinates, calculated on the perpendicular bisector of the baseline shown in Figure 9.14, depends on

$$\sigma_x = \frac{c\sigma_t \sqrt{\left(\frac{b}{2}\right)^2 + D^2}}{D} \tag{9.61}$$

and

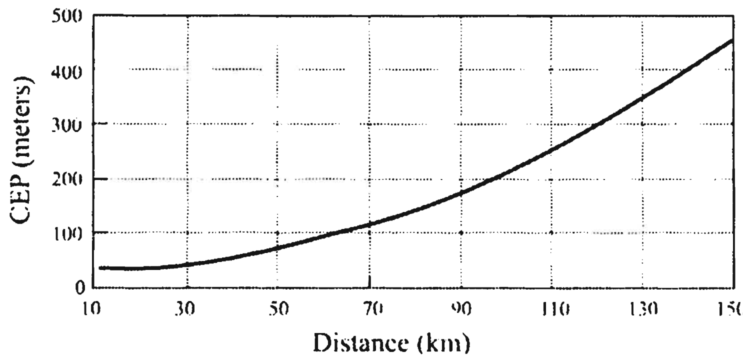


Figure 9.14 CEP as a function of distance from the sensor baseline on the perpendicular bisector.

$$\sigma_y = \frac{\lambda\sigma_f\left[\left(\frac{b}{2}\right)^2 + D^2\right]^{3/2}}{vbD} \tag{9.62}$$

where λ is the wavelength, v is the velocity of the EW systems, b is the baseline length, and D is the distance from the baseline.

The standard deviations of the fix calculation form the major and minor axes of the elliptical error probable contour. As given previously, an approximate (to within 10%) CEP contour can be determined from these two values as

$$\text{CEP} = 0.75\sqrt{\sigma_x^2 + \sigma_y^2} \tag{9.63}$$

To get a sense of the geolocation accuracy possible with τ and $\dot{\tau}$ processing, consider the following example. Suppose that the target transmits enough power so that an SNR of 10 dB is received at each of two airborne EW systems. Further suppose that the transmitted frequency is 50 MHz, the signal bandwidth is 25 kHz, the integration time is 0.3 second, the distance between the systems is 50 km, and the velocity of the airborne systems is 32 m/s laterally to the target. The geolocation accuracy possible on the perpendicular bisector to the baseline is shown in Figure 9.14.

As can be seen, geolocation accuracy of less than 500m is possible with these parameters out to a range of 150 km and more (assuming a constant SNR of 10 dB). Note, however, that these results are along the perpendicular bisector of the baseline. Off the baseline the GDOP decreases the accuracy as discussed above.

9.5 Time of Arrival

Geolocation of a target based on measuring the τ and $\dot{\tau}$ of an EM wave emanating from the target was presented in Section 9.4. Presented in this section is an algorithm for computing the geolocation of a target based on the measurement of the TOAs of the signal wave front at four (or more) EW systems. This derivation follows that in [31] closely. Let the location of the target be denoted by $r_\text{T} = (x_\text{T}, y_\text{T}, z_\text{T})$ and the location of the systems be denoted as $r_i = (x_i, y_i, z_i)$, $i = 1, 2, 3, 4$. Further let the time of transmission of a signal be denoted by t_0, and t_i, $i = 1, 2, 3, 4$ be the arrival time of the signal at each of the receiving systems. The Euclidean distances between the systems and the target are given by

$$d_i = \left\| r_i - r_\text{T} \right\| = \sqrt{(x_i - x_\text{T})^2 + (y_i - y_\text{T})^2 + (z_i - z_\text{T})^2} \tag{9.64}$$

and d_i is related to the arrival times t_i by

$$d_i = c\left(t_i - t_0\right) \tag{9.65}$$

where c is the speed of light. Therefore, by equating these expressions and squaring, we obtain

$$\begin{aligned} c^2t_1^2 - 2c^2t_1t_0 + c^2t_0^2 - x_1^2 + 2x_1x_\text{T} - x_\text{T}^2 - y_1^2 + {} \\ 2y_1y_\text{T} - y_\text{T}^2 - z_1^2 + 2z_1z_\text{T} - z_\text{T}^2 = 0 \end{aligned} \tag{9.66}$$

$$\begin{aligned} c^2t_2^2 - 2c^2t_2t_0 + c^2t_0^2 - x_2^2 + 2x_2x_\text{T} - x_\text{T}^2 - y_2^2 + {} \\ 2y_2y_\text{T} - y_\text{T}^2 - z_2^2 + 2z_2z_\text{T} - z_\text{T}^2 = 0 \end{aligned} \tag{9.67}$$

$$\begin{aligned} c^2t_3^2 - 2c^2t_3t_0 + c^2t_0^2 - x_3^2 + 2x_3x_\text{T} - x_\text{T}^2 - y_3^2 + {} \\ 2y_3y_\text{T} - y_\text{T}^2 - z_3^2 + 2z_3z_\text{T} - z_\text{T}^2 = 0 \end{aligned} \tag{9.68}$$

$$\begin{aligned} c^2t_4^2 - 2c^2t_4t_0 + c^2t_0^2 - x_4^2 + 2x_4x_\text{T} - x_\text{T}^2 - y_4^2 + {} \\ 2y_4y_\text{T} - y_\text{T}^2 - z_4^2 + 2z_4z_\text{T} - z_\text{T}^2 = 0 \end{aligned} \tag{9.69}$$

By subtracting these equations two at a time, t^2 can be eliminated from all of them, thereby making them a linear set of equations in t_0. This must be done, however, so that the resulting equations are independent.

Subtracting (9.36b) from (9.36a), then (9.36c) from (9.36b), and finally (9.36d) from (9.36c) yields

$$\begin{aligned} &c^2\left(t_1^2 - t_2^2\right) - 2c^2\left(t_1 - t_2\right)t_0 - \left(x_1^2 - x_2^2\right) + 2\left(x_1 - x_2\right)x_T \\ &\qquad - \left(y_1^2 - y_2^2\right) + 2\left(y_1 - y_2\right)y_T - \left(z_1^2 - z_2^2\right) + 2\left(z_1 - z_2\right)z_T = 0 \end{aligned} \tag{9.70}$$

$$\begin{aligned} &c^2\left(t_2^2 - t_3^2\right) - 2c^2\left(t_2 - t_3\right)t_0 - \left(x_2^2 - x_3^2\right) + 2\left(x_2 - x_3\right)x_T \\ &\qquad - \left(y_2^2 - y_3^2\right) + 2\left(y_2 - y_3\right)y_T - \left(z_2^2 - z_3^2\right) + 2\left(z_2 - z_3\right)z_T = 0 \end{aligned} \tag{9.71}$$

$$\begin{aligned} &c^2\left(t_3^2 - t_4^2\right) - 2c^2\left(t_3 - t_4\right)t_0 - \left(x_3^2 - x_4^2\right) + 2\left(x_3 - x_4\right)x_T \\ &\qquad - \left(y_3^2 - y_4^2\right) + 2\left(y_3 - y_4\right)y_T - \left(z_3^2 - z_4^2\right) + 2\left(z_3 - z_4\right)z_T = 0 \end{aligned} \tag{9.72}$$

Rearranging terms

$$\begin{aligned} &\left(x_1 - x_2\right)x_T + \left(y_1 - y_2\right)y_T + \left(z_1 - z_2\right)z_T = \\ &\qquad c^2\left(t_1 - t_2\right)t_0 - \frac{1}{2}c^2\left(t_1^2 - t_2^2\right) + \frac{1}{2}\left(x_1^2 - x_2^2 + y_1^2 - y_2^2 + z_1^2 - z_2^2\right) \end{aligned} \tag{9.73}$$

$$\begin{aligned} &\left(x_2 - x_3\right)x_T + \left(y_2 - y_3\right)y_T + \left(z_2 - z_3\right)z_T = \\ &\qquad c^2\left(t_2 - t_3\right)t_0 - \frac{1}{2}c^2\left(t_2^2 - t_3^2\right) + \frac{1}{2}\left(x_2^2 - x_3^2 + y_2^2 - y_3^2 + z_2^2 - z_3^2\right) \end{aligned} \tag{9.74}$$

$$\begin{aligned} &\left(x_3 - x_4\right)x_T + \left(y_3 - y_4\right)y_T + \left(z_3 - z_4\right)z_T = \\ &\qquad c^2\left(t_3 - t_4\right)t_0 - \frac{1}{2}c^2\left(t_3^2 - t_4^2\right) + \frac{1}{2}\left(x_3^2 - x_4^2 + y_3^2 - y_4^2 + z_3^2 - z_4^2\right) \end{aligned} \tag{9.75}$$

Recognizing that

$$\begin{aligned} \left\|\vec{r}_1\right\|^2 &= x_1^2 + y_1^2 + z_1^2 \\ \left\|\vec{r}_2\right\|^2 &= x_2^2 + y_2^2 + z_2^2 \\ \left\|\vec{r}_3\right\|^2 &= x_3^2 + y_3^2 + z_3^2 \\ \left\|\vec{r}_4\right\|^2 &= x_4^2 + y_4^2 + z_4^2 \end{aligned} \tag{9.76}$$

these equations can be written in matrix form as

$$\mathbf{A}\vec{r}_T = c^2\vec{u}t_0 + \vec{s} \tag{9.77}$$

where

$$\mathbf{A} = \begin{bmatrix} x_1 - x_2 & y_1 - y_2 & z_1 - z_2 \\ x_2 - x_3 & y_2 - y_3 & z_2 - z_3 \\ x_3 - x_4 & y_3 - y_4 & z_3 - z_4 \end{bmatrix}$$

$$\vec{u} = \begin{bmatrix} t_1 - t_2 \\ t_2 - t_3 \\ t_3 - t_4 \end{bmatrix}$$

$$\vec{s} = \frac{1}{2}\begin{bmatrix} \|\vec{r}_1\|^2 - \|\vec{r}_2\|^2 - c^2\left(t_1^2 - t_2^2\right) \\ \|\vec{r}_2\|^2 - \|\vec{r}_3\|^2 - c^2\left(t_2^2 - t_3^2\right) \\ \|\vec{r}_3\|^2 - \|\vec{r}_4\|^2 - c^2\left(t_3^2 - t_4^2\right) \end{bmatrix}$$

The solution to this equation is given by

$$\vec{r}_\text{T} = c^2\mathbf{A}^{-1}\vec{u}t_0 + \mathbf{A}^{-1}\vec{s} \tag{9.78}$$

assuming that $\mathbf{A}^{-1}$ exists, which is true here. Let

$$\begin{aligned} \vec{v} &= \mathbf{A}^{-1}\vec{u} \\ \vec{w} &= \mathbf{A}^{-1}\vec{s} \end{aligned} \tag{9.79}$$

Substituting this solution into one of the equations at (9.34) yields

$$d_1^2 = \|\vec{r}_T - \vec{r}_1\|^2 = c^2\left(t_1 - t_0\right)^2 \tag{9.80}$$

$$\|\vec{v}t + \vec{w} - \vec{r}_1\|^2 = c^2 t_1^2 - 2c^2 t_1 t_0 + c^2 t_0^2 \tag{9.81}$$

$$\|\vec{v}t + (\vec{w} - \vec{r}_1)\|^2 - c^2 t_0^2 + 2c^2 t_1 t_0 - c^2 t_1^2 = 0 \tag{9.82}$$

Let

$$\vec{v} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} \tag{9.83}$$

$$\vec{w}-\vec{r}_1 = \begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix} \tag{9.84}$$

then

$$\begin{aligned}
\left\|\vec{v}t+(\vec{w}-\vec{r}_1)\right\|^2 &= \left(v_1 t_0 - q_1\right)^2 + \left(v_2 t_0 - q_2\right)^2 + \left(v_3 t_0 - q_3\right)^2 \\
&= v_1^2 t_0^2 - 2v_1 q_1 t_0 + q_1^2 + v_2^2 t_0^2 - 2v_2 q_2 t_0 + q_2^2 + v_3^2 t_0^2 - 2v_3 q_3 t_0 + q_3^2 \\
&= \left(v_1^2 + v_2^2 + v_3^2\right)t_0^2 - 2\left(v_1 q_1 + v_2 q_2 + v_3 q_3\right)t_0 + \left(q_1^2 + q_2^2 + q_3^2\right) \\
&= \left\|\vec{v}\right\|^2 t_0^2 - 2\vec{v}^{\mathrm{T}}\left(\vec{w}-\vec{r}_1\right)t_0 + \left\|\vec{w}-\vec{r}_1\right\|^2
\end{aligned} \tag{9.85}$$

Therefore,

$$\begin{aligned}
\left\|\vec{v}\right\|^2 t_0^2 - 2\vec{v}^{\mathrm{T}}\left(\vec{w}-\vec{r}_1\right)t_0 + \left\|\vec{w}-\vec{r}_1\right\|^2 - c^2 t_0^2 + 2c^2 t_1 t_0 - c^2 t_1^2 = 0 \\
\left(\left\|\vec{v}\right\|^2 - c^2\right)t_0^2 + 2\left[c^2 t_1 - \vec{v}^{\mathrm{T}}\left(\vec{w}-\vec{r}_1\right)\right]t_0 + \left[\left\|\vec{w}-\vec{r}_1\right\|^2 - c^2 t_1^2\right] = 0
\end{aligned} \tag{9.86}$$

Equation (9.86) can be solved for t_0, the time that the signal was transmitted. Two answers ensue, corresponding to the two roots of the polynomial. Other information is necessary, then, to ascertain which of the two answers is correct. Once t_0 is determined, it can then be substituted into (9.42) to determine $\vec{r}_{\mathrm{T}}$. This, of course, works for any of the original equations, not just $i = 1$.

9.6 TDOA Asymptote Emitter Localization

Rather than estimating the location of the target based on the intersection of TDOA hyperbolas, Dogancay proposed to use the intersection of the asymptotic extensions of the hyperbolas as the location estimate [49]. Such a technique is considerably faster than finding the intersection of the hyperbolas and provides reasonable accuracy.

9.7 Multiple Targets

An important problem in any realistic employment of TDOA or other types of geolocation techniques is dealing with multiple targets. In most practical military applications, the frequency spectrum is extremely crowded—consisting of both friendly as well as target signals. The target association problem was addressed by Sathyan et al. [50].

9.8 Concluding Remarks

Hyperbolic position-fixing techniques based on measuring the TOA, TDOA, or differential Doppler at two or more widely separated platforms are presented in this chapter. These techniques have been shown in practice to yield better accuracy than the direction-finding techniques presented in Chapter 9. The possible accuracies from such systems are illustrated in Figure 9.12. In general, however, longer integration times are required than for the DF approaches.

In addition to processing time, typically considerably more signal processing is required in CAF processing to compute the ambiguity surfaces. High-speed signal processors are required as the surfaces typically consist of considerable data points.

Lastly, as would be expected, accurate knowledge of the platform parameters is required to exploit the accurate PF capabilities of the quadratic technologies. These parameters are the three-dimensional positions, velocities, and accelerations.

References

[1] Chan, Y. T., and K. C. Ho, "An Efficient Closed-Form Localization Solution from Time Difference of Arrival Measurements," *Proceedings IEEE MILCOM*, 1994, pp. II-393–II-396.

[2] Ho, K. C., and Y. T. Chan, "Solution and Performance Analysis of Geolocation by TDOA," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 29, No. 4, October 1993, pp. 1311–1322.

[3] Stein, S., "Algorithms for Ambiguity Function Processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-29, No. 3, June 1981, pp. 588–599.

[4] Ho, K. C., and Y. T. Chan, "Geolocation of a Known Altitude Object from TDOA and FDOA Measurements," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 33, No. 3, July 1997, pp. 770–783.

[5] Chan, Y. T., and K. C. Ho, "A Simple and Efficient Estimator for Hyperbolic Location," *IEEE Transactions on Signal Processing*, Vol. 42, No. 8, August 1994, pp. 1905–1915.

[6] Bard, J. D., F. M. Ham, and W. L. Jones, "An Algebraic Solution to the Time Difference of Arrival Equations," *Proceedings IEEE MILCOM*, 1996, pp. 313–319.

[7] Schau, H. C., and A. Z. Robinson, "Passive Source Localization Employing Intersecting Spherical Surfaces from Time-of-Arrival Differences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-35, No. 8, August 1987, pp. 1223–1225.

[8] Quazi, A. H., "An Overview on the Time Delay Estimate in Active and Passive Systems for Target Localization," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-29, No. 3, June 1981, pp. 527–533.

[9] Campbell, L. L., "Asymptotics of Performance of Estimators of Arrival Time," *ISIT*, 1998.

[10] Gardner, W. A., and C. K. Chen, "Interference-Tolerant Time-Difference-of-Arrival Estimation for Modulated Signals," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 36, No. 9, September 1988, pp. 1385–1395.

[11] Rusu, P., "A TOA-FOA Approach to Find the Position and Velocity of RF Emitters," Applied Research Laboratories, The University of Texas at Austin.

[12] Yost, G. P., and S. Panchapakesan, "Automatic Location Identification Using a Hybrid Technique," *Proceedings Vehicular Technology Conference*, 1998, pp. 264–267.

[13] Yuan, Y. X., G. C. Carter, and J. E. Salt, "Correlation Among Time Difference of Arrival Estimators and Its Effect on Localization in a Multipath Environment," *Proceedings IEEE MILCOM*, 1995, pp. 3163–3166.

[14] Chestnut, P. C., "Emitter Location Accuracy Using TDOA and Differential Doppler," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. AES-18, No. 2, March 1982, pp. 214–218.

[15] Belanger, S. P., "An EM Algorithm for Multisensor TDOA/DD Estimation in a Multipath Propagation Environment," *Proceedings IEEE MILCOM*, 1996, pp. 3110–3120.

[16] Schmidt, R., "Least Squares Range Difference Location," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 32, No. 1, January 1996, pp. 234–242.

[17] Shin, D. C., and C. L. Nikias, "Complex Ambiguity Function Based on Fourth-Order Statistics for Joint Estimation of Frequency-Delay and Time-Delay of Arrival," *Proceedings IEEE MILCOM*, 1993, pp. 461–465.

[18] Smith, J. O., and J. S. Abel, "Closed-Form Least-Squares Source Location Estimation from Range-Difference Measurements," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-35, No. 2, December 1987, pp. 1661–1669.

[19] Fang, B. T., "Simple Solutions for Hyperbolic and Related Position Fixes," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 26, No. 5, September 1990, pp. 748–753.

[20] Wiley, R. G., *Electronic Intelligence: The Interception of Radar Signals*, Dedham, MA: Artech House, 1985, pp. 116–121.

[21] Bard, J. D., and F. M. Ham, "Time Difference of Arrival Dilution of Precision and Applications," *IEEE Transactions on Signal Processing*, Vol. 47, No. 2, February 1999, pp. 521–523.

[22] Whalen, A. D., *Detection of Signals in Noise*, New York: Academic Press, 1971, p. 339.

[23] Piersol, A., "Time Delay Estimation Using Phase Data," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-29, No. 3, June 1981, pp. 471–477.

[24] Rodriguez, M. A., R. H. Williams, and T. J. Carlow, "Signal Delay and Waveform Estimation Using Unwrapped Phase Averaging," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-29, No. 3, June 1981, pp. 508–513.

[25] Saarnisaari, H., "ML Time Delay Estimation in a Multipath Channel," *Proceedings IEEE 1996 4th International Symposium on Spread Spectrum Techniques and Applications*, Vol. 3, September 22–25 1996, pp. 1001–1011.

[26] Owsley, N. L., and G. R. Swope, "Time Delay Estimation in a Sensor Array," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-29, No. 3, June 1981, pp. 519–523.

[27] Knapp, C. H., and G. C. Carter, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-24, August 1976, pp. 320–327.

[28] Wuu, C-Y, and A. E. Pearson, "Time-Delay Estimation Involving Received Signals," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-32, No. 4, August 1984, pp. 828–835.

[29] Azania, M., and D. Hertz, "Time Delay Estimation by Generalized Cross Correlation Methods," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-32, No. 2, April 1984, pp. 280–285.

[30] Krolik, J., M. Eizenman, and S. Pasupathy, "Time Delay Estimation of Signals with Uncertain Spectra," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 36, No. 12, December 1988, pp. 1801–1811.

[31] Rusu, P., "The Equivalence of TOA and TDOA RF Transmitter Location," Applied Research Laboratories, The University of Texas at Austin, Austin, TX.

[32] Cusani, R., "Performance of Fast Time Delay Estimators," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 37, No. 5, May 1989, pp.757–759.

[33] Nikias, C. L., and R. Pan, "Time Delay Estimation in Unknown Gaussian Spatially Correlated Noise," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 36, No. 11, November 1988, pp. 1706–1714.

[34] Rusu, P., and L. C. Giulianelli, "Covariance of Time-Difference of Arrival Residuals Among Pairs of Observers for RF Emitter Location Systems," The Applied Physics Laboratories, The University of Texas at Austin.

[35] Yuan, Y. X., G. C. Carter, and J. E. Salt, "Correlation among Time Difference of Arrival Estimators and its Effect on Localization in a Multipath Environment," *Proceedings IEEE 1995 International Conference on Acoustics, Speech, and Signal Processing*, ICASSP-95, Vol. 5, pp. 3163–3166.

[36] Bard, J. D., and F. M. Ham, "Time Difference of Arrival Dilution of Precisions and Applications," *IEEE Transactions on Signal Processing*, Vol. 47, No. 2, February 1999, pp. 521–523.

[37] Bard, J. D., F. Ham, and W. L. Jones, "An Algebraic Solution to the Time Difference of Arrival Equations," *Proceedings IEEE Southeastern Conference*, Tampa, Fl, April 11–14, 1996, pp. 313–319.

[38] Torrerri, D. J., "Statistical Theory of Passive Location Systems," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. AES-20, No. 2, March 1984, pp. 183–198.

[39] Weiss, A. J., and E. Weinstein, "Fundamental Limitations in Passive Time Delay Estimation, Part I: Narrowband Systems," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-31, No. 2, April 1983, pp. 472–485.

[40] Weinstein, E., and A. J. Weiss, "Fundamental Limitations in Passive Time Delay Estimation, Part II: Wide-Band Systems," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-32, No. 5, October 1984, pp. 1064–1077.

[41] Weiss, A. J., "Composite Bound on Arrival Time Estimation Errors," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. AES-22, No. 6, November 1986, pp. 751–756.

[42] Weiss, A. J., and Z. Stein, "Optimal Below Threshold Delay Estimation for Radio Signals," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. AES-23, No. 6, November 1987, pp. 726–730.

[43] Weiss, A. J., "Bounds on Time-Delay Estimation for Monochromatic Signals," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. AES-23, No. 6, November 1987, pp. 798–808.

[44] Chow, S-K., and P. M. Shelties, "Delay Estimation Using Narrow-Band Processes," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-29, No. 3, June 1981, pp. 478–484.

[45] Campbell, L. L., "Asymptotics of Performance of Estimators of Arrival Time," *ISIT* 1998, Cambridge, MA, August 16–21.

[46] Chow, S. K., and P. M. Schultheiss, "Delay Estimation Using Narrow-Band Processes," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-29, No. 3, June 1981, pp. 478–484.

[47] Stein, S., "Algorithms for Ambiguity Function Processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-29, No. 3, June 1981, pp. 588–599.

[48] Chestnut, P. C., "Emitter Location Accuracy Using TDOA and Differential Doppler," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. AES-18, No. 2, March 1982, pp. 214–218.

[49] Dogancay, K., "Emitter Localization Using Clustering-Based Bearing Association," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 41, No. 2, April 2005, pp. 525–536.

[50] Sathyan, T., A. Sinha, and T. Kirubarajan, "Passive Geolocation and Tracking of an Unknown Number of Emitters," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 42, No. 2, April 2006, pp. 740–749.

Part II – Performance

Chapter 10

Early-Entry Organic Electronic Support

10.1 Introduction

In any kind of hostile activity, the U.S. military has the requirement to collect battlefield combat information about the adversary's disposition and intent. One way to assist in addressing this requirement is to intercept the tactical communications associated with the adversary using ground-based and airborne collection assets. Early-entry forces are constrained in terms of organic support available due primarily to airlift availability. Intelligence support to these forces may be furnished from higher echelons and, in particular, national assets. In some cases, however, it may be desirable due to access and other reasons to provide ES with organic assets. The collection capabilities that could be provided this way would naturally be very limited. Lightweight, manpack type of equipment would be all that could reasonably be provided [1–8].

This chapter presents how well a ground-based communications intercept system would perform against a threat communications environment that might be encountered by early-entry forces. It is based on a simulation that compares, in an operational and engineering sense, the values of various system tradeoffs that can affect performance. For example, does the requirement to prepare preformatted reporting messages, based on an intercept, significantly degrade the system output?

There were two independent models used in the simulation: a target environment model and a collection system model. The model of the signal environment did not represent any particular part of the world or any particular hostile threat. The target frequencies were randomly selected from the low-VHF military frequency range and were randomly placed geographically within an area, which was 30 km square. The model of the collection system did not represent any particular existing system but did closely and with a fair amount of detail model both the operational and systemic steps necessary to perform the communications intercept mission.

10.2 Target Model

The target environment consisted of a selectable number of simplex, clear text, VHF, PTT, ground, stationary communication nets. The tactical communications taking place within the first hour after the intercept systems were put into operation were simulated. No a priori knowledge of the target environment (such as frequencies) was assumed (which is a pessimistic assumption). The number of threat communication nets was a simulation variable: 100, 200, and 300 nets. Each net consisted of five nodes. This would approximate a brigade-sized adversary force.

In any live case, communications transpiring over tactical communication nets could take several forms. For the purposes of modeling, a particular communication structure was simulated. In each case of a communication exchange, net member one, called the NCS, transmitted first. The length of this transmission was (almost) a Gaussian random amount of time with an average of 5 seconds, a standard deviation of 1.67 seconds, and a minimum time of 1 second. This transmission was followed by a period of silence, the duration of which was randomly selected with an average value of 3 seconds, a standard deviation of 1.33 seconds, and a minimum of 1 second. Net member two then transmitted for a random time with the same statistics as the NCS. This transmission was followed by a period of silence with the same statistics as above, followed by the NCS transmitting to station three and station three responding. After that, the NCS transmitted to station four and station four responded. Finally, the NCS transmitted to station five and station five responded. Each net communicated in this fashion repetitively, the timing of which was (almost) a Gaussian random variable called the *net time*. The mean of this variable was varied between 150 seconds and 450 seconds, while the standard deviation was varied between 50 and 150 seconds. This process is shown schematically in Figure 10.1.

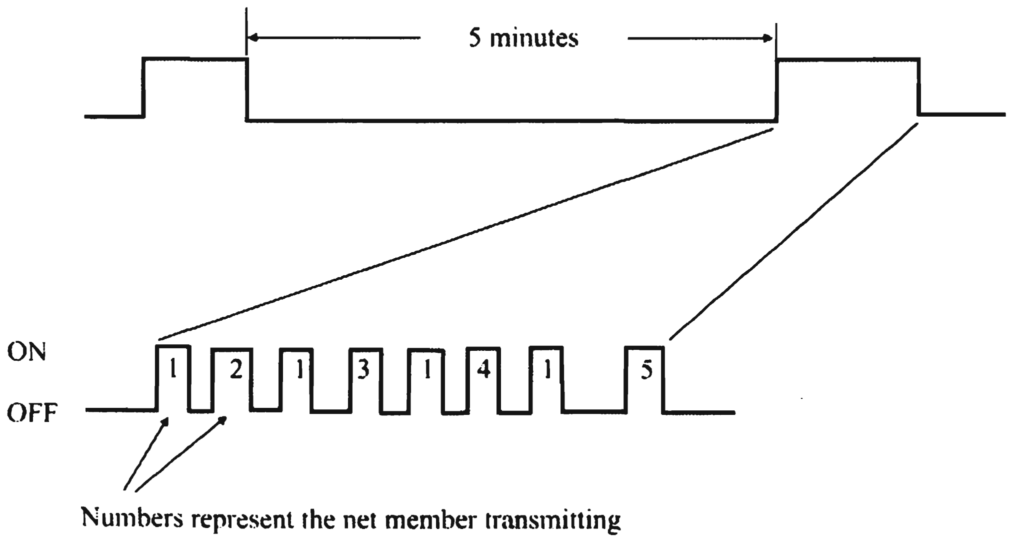


Figure 10.1 Target communication link timing.

The target nets were placed randomly within a 30 km × 30 km area. The node separations on any given net were tactically reasonable. Thus, the emitters were not concentrated in any particular area. This may not be as accurate as it could have been since normally one would expect a greater concentration of emitters closer to the *forward line of troops* (FLOT), if there is one, than at the rear. The parameter for the Longley-Rice propagation model that was used presumed that the transmitters were randomly sited. (This has to do with how carefully the transmitters were sited for communication purposes.) Throughout the duration of the simulation, the targets were assumed to be stationary.

The targets were assumed to all have an effective radiated power of 20W with vertical polarization and the transmitter antenna heights were 5m. Signal losses with distance were computed using the Longley-Rice signal propagation model [9] using values of terrain variables that are considered average for typical ground conditions found in many parts of the world, with gently rolling terrain andsoil that is not dry, but not too wet. Specifically, it was an equatorial temperate climate with a Δh value of 100m. This model has been used extensively to approximate signal losses that occur close to the Earth, such as the case here.

The principal quality that constrains signal intercepts is the SNR of the transmission at the receive site. Therefore, so that an SNR could be calculated, a noise model was included in the simulation. The noise model used is based on the CCIR report described in Chapter 3, which includes a typical rural noise environment. The amount of noise present at the frequency of transmission was added to the signal received at the receiving system and if the SNR did not exceed a specified value, which depended on whether intercept or DF was being performed, the signal was considered as either detectable or not detectable.

10.3 Intercept System Model

The intercept systems were comprised of three receiving systems with DF capability. Each system had one operator, typical of manpack configurations. These receivers had the capability to step between frequencies (from a preprogrammed set) and scan between two frequencies set by the operators. If energy were detected at a channel (frequency), then the receiver dwelled there until the collection operation was concluded (defined later) at which time the next frequency was selected. The search mode used in the simulation consisted of starting with a frequency band assigned to each operator, the particular band being selected by dividing the 30–90 MHz tactical VHF, PTT frequency band into equal increments. Each segment was then assigned to a different operator [10–12].

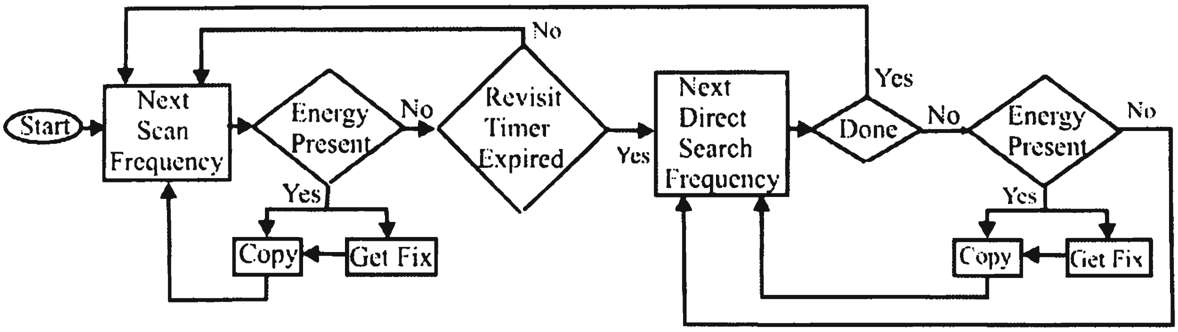


Figure 10.2 Flow diagram of the overall operation of the collection systems.

When the simulation started, each VHF receiver began scanning from its lowest limit and stepped in one-half *IF bandwidth* (IFBW) steps up to its maximum limit, the IFBW being a parameter in the simulation. The system dwelled 25 ms at any channel unless energy was detected there. When energy was detected at a particular frequency, that frequency was entered into a directed search list, and the operator prosecuted (copied) that particular frequency. Figure 10.2 shows an overview of the operation of the intercept systems.

Because communications are usually between (at least) two people, it is a good idea to continue to dwell for some time on a channel after a transmission has ceased, in order to hear the follow-on communication (if there is one). That capability was included in the simulation; the length of the dwell continuation is called the *bridge time*. This timer was set at 15 seconds. The bridge timer started when the signal dropped below the intercept SNR, which was a simulation parameter. For all of the simulations, a 6 dB SNR was required for the receiver to initially stop scanning and, once dwelling on a target, a 3 dB SNR was necessary to remain there. If a 3 dB SNR was not present, then the bridge timer continued running. Once the bridge timer expired, then a period of postprocessing time occurred. The length of this postprocessing time was a variable set at 15, 100, and 200 seconds. After the postprocessing, the frequency scanning continued.

Once a channel was found that contained a target signal, that channel became a priority frequency, which should be revisited more often than the other channels. In the general search mode described above it was indicated that these frequencies were saved on a list called the *directed search* list. The length of time spent in the general search mode before scanning was interrupted and the directed search frequencies were examined for activity, which is called the *revisit time*. The revisit time for this simulation was 7 seconds.

The intercept systems were netted together so that target fixes could be obtained. The DF integration time (the amount of time necessary to obtain an LOB) was set at 2 seconds and the net cycle time (the total time required to obtain a fix) was 8 seconds. If the DF net was busy when the operator requested a fix, then the request went into a queue of length one. If the queue was already full, then the request was denied. Any operator in any system could request a fix but the queue length was one for the entire net. In order for a system to obtain an LOB that may be used in a subsequent fix calculation, an SNR of –6 dB or more must have been present. The simulation model included Gaussian random errors added to the LOBs yielding an effective LOB accuracy of 0° average and 3° rms error. The fix algorithm used was that due to Brown described in Section 10.3.3.

No friendly signals interfered with the intercept process because the capability to automatically recognize the 150-Hz squelch tone present on NATO radios was assumed present. Therefore, modeling of the friendly radio frequency environment was not necessary.

Each operator was assumed to have a computer on which he or she could compose preformatted messages describing intercept operations and report them out of the intercept system to an analysis center. Modeling of the analysis center was not part of the simulation. The time it took to prepare these messages was a simulation variable called the *postprocessing time*. When it is set to 15 seconds (its lowest value), this approximates the situation in which the operator terminal operates autonomously in that target locations and frequencies are automatically sent to the analysis center without the necessity of the operator getting involved. This was the mode where the most target searching could be performed. The postprocessing times were 15 seconds, 100 seconds, and 200 seconds to illustrate the effects of having the operator prepare brief, but finite reports. Such reports would typically be, in addition to the target locations and frequencies, a brief synopsis of the transmissions intercepted on preformatted message formats. Such information, for example, might include which of the net nodes were the NCS. After the postprocessing time, the operator and equipment were then free to continue intercept operations.

The intercept systems were placed on the friendly side of the assumed FLOT at coordinates (–5, 7.5), (–8, 15), and (–5, 22.5). [The FLOT is assumed to be a vertical line running from coordinates (0, 0) to (0, 30).] All distances are in kilometers. The antenna height for the intercept systems was 7m.

In addition to the external noise modeled as indicated above, the RF receiving components in the collection system added thermal noise. This phenomenon is manifest in the equivalent system *noise figure* (NF) discussed in Chapter 5. The NF was a parameter in the simulation and was set at 10, 15, and 20 dB. The amount of external noise present and the NF together were used to set the noise values used in the SNR calculations. Although not an operational parameter that can be controlled, the NF is one of the most important design tradeoff parameters that, among other variables, set system sensitivity.

The IFBW(and for these discussions also called the search bandwidth) was also a model parameter that was varied between 10 kHz and 250 kHz. The wider the search bandwidth, the faster the scanning process proceeds; however, the receive noise varies directly with the bandwidth as discussed in Chapter 5. Therefore, a tradeoff is typically necessary between search time and search bandwidth.

After a signal goes away, the receiver continued to dwell at the tuned frequency for a duration known as the *dwell time.* During the period immediately after the termination of a signal was the time when the system looked for another transmission from other members on the same net.

The antenna height of receive systems, in addition to the system sensitivity, largely determines how far into the target area the collection systems can hear. For this simulation this height was 7m. As mentioned, the system sensitivity depends on the IF bandwidth of the receivers, which was set at 10 kHz, 50 kHz, and 250 kHz.

The wider IFBW then searching the spectrum occurred at a more rapid pace, thereby increasing the probability of detecting a signal earlier in a transmission. However, the wider the bandwidth, the more noise was allowed into the passband and therefore the SNR was reduced, thus making distant targets harder to detect. Wider bandwidths also overlap several VHF frequency channels and thereby increase the possibility of cochannel interference (more than one signal in the passband, making listening to either one harder). The system sensitivities are shown in Table 10.1.

The antennas used in the simulation were "typical" antennas that might be used with ground-based communications intercept systems. They were omnidirectional in the horizontal plane and had gains that were about –20 dB at the low end of the band rising to about –3 dB at the upper end of the frequencies used in the simulation. Of course, the height of the receiving (and transmitting, for that matter) antennas, to a large extent, determines the amount of signal power that can be received. On the other hand, the amount of noise present at the receiving site does not depend on the receiver antenna height.

Table 10.1 Approximate System Sensitivity as a Function of IFBW and Noise Figure (in dBm)

	Noise Figure (dB)		
IFBW	**10**	**15**	**20**
10 kHz	–124	–119	–114
50 kHz	–117	–112	–107
250 kHz	–110	–105	–100

10.4 Simulation Results

The parameters that were varied in this analysis were: (1) the number of target nets, (2) the search bandwidth of the receivers, (3) the system noise figure, (4) post-processing time, and (5) the duration of the simulation. A *target* herein refers to a particular communication node or radio. A *target net* or just *net* is a set of targets that are operating on the same frequency and occasionally communicate with one another. Each net had a *net control station* that communicated more than the others; normally it would be associated with a unit commander. A net was considered *detected* if one or more of the transmissions associated with the net were collected. A *target* was considered *detected* if one or more transmissions were collected from that target. Since it is routine practice to transmit one's own call sign at the beginning of a communiqué, a target was considered *identified* if one or more transmissions were collected from that target and the collection started within the first 2 seconds of the transmission. The a posteriori probabilities, probability of detection of a target transmitter (at least once), denoted P_{det}, probability of identification of a target transmitter (at least once), denoted P_{id}, probability of location of a target transmitter (at least once), denoted P_{loc}, probability of identification of a net's control station, denoted $P_{\text{id,NCS}}$, and probability of location of a net's control station, denoted $P_{\text{loc,NCS}}$, are the statistics that were collected and analyzed.

There were three reasons why a fix may not have been obtained on a target transmitter: (1) there was too little time left when the data link was acquired (the signal went away before the DF acquisition started), (2) there were too few LOBs available (< 2) because of too weak a signal at too many collection sites, and (3) the DF queue was full when the operator requested use of the DF assets. All else being equal, that is, all of the targets could be "heard" by all of the stations, if the DF assets were busy all the time, then 450 fixes per hour could be obtained.

It was possible to obtain a fix on the wrong target because the fix process took a certain period of time. When the operator requested a fix and the DF assets were busy, the request may have gone into the queue of length one. When the DF assets became available, the DF request was honored but by that time a different net member may have become active, in which case the fix was on the wrong target. Since a target was identified by its frequency of operation, even though the specific target was in error, the fix would have been on another member of the same net, however. This could cause confusion as to whether the target located was the NCS or not, but the results in tactical scenarios is not useless data in that case. The total number of wrong fixes indicates the number of occasions of this happening.

Ten passes through the simulation were performed using different random seeds for each condition of the simulation variables. The mean values of the results were calculated and used in the analysis to follow.

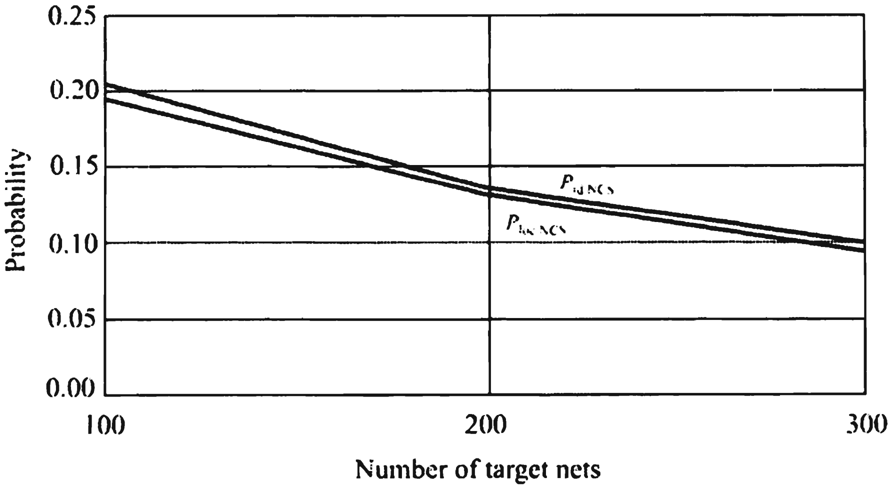


Figure 10.3 NCS collection results versus the number of target nets.

10.4.1 Performance Versus the Number of Target Nets

The results as the number of target nets was varied and are shown in Figures 10.3 and 10.4. The ability to find net control stations is shown in Figure 10.3. As expected, the number found decreases as the number of nets is increased. This is because (1) the operators of the intercept systems on average were busier because there were more targets with which to contend and (2) more of the target nets were probably out of range, which decreases the percentage results. Since, on average, a transmission would be detected halfway through, and considering that the dwell time was 15 seconds, with little postprocessing time an operator would spend approximately 46 seconds per intercept. That means that one operator could process about 78 targets per

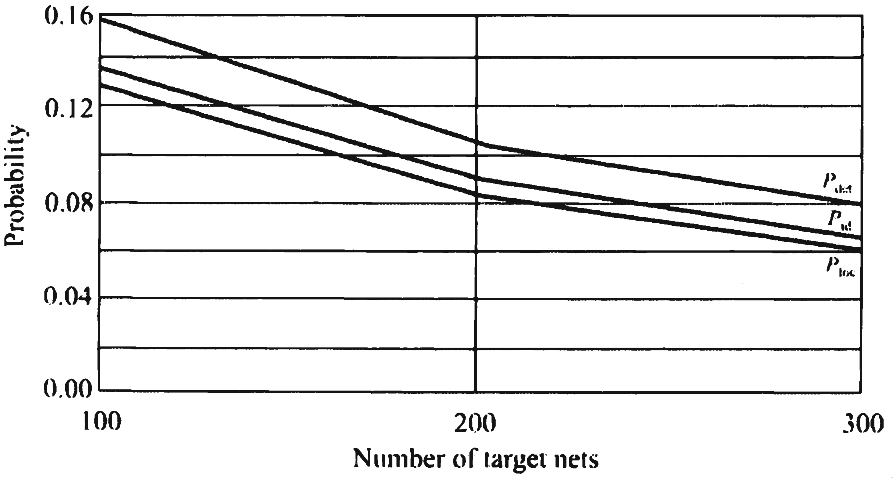


Figure 10.4 General collection performance versus the number of target nets.

hour if that operator were employed full time. Note that some of these targets need not be unique; some could be duplicates. Therefore, under these conditions, a maximum of 234 targets could be prosecuted. Ignoring hearability considerations, then, a maximum of 234 nets could be detected. Due to the nature of the communication on the target nets shown in Figure 10.1, $P_{\text{id,NCS}}$ would be this same level of probability since on the average a net cycle would be encountered in the middle. Also due to this, the latter net members would stand a better chance of being detected than the earlier ones, yielding an expected P_{id} about one half that of $P_{\text{id,NCS}}$.

The above analysis ignored any limiting effects of hearability. The receiver antenna heights were only 7m. While this is practical from a deployment point of view, it does not facilitate much range from the intercept sites. Of course, if available, mountaintops are useful to extend this range. Such mountaintops are rarely the lone province of EW systems, however, and must be shared with other facilities that like such high ground. Nevertheless, there was not much range from such short antennas. Certainly not enough to extend to the 30 km depth of the target space. Hence, as more targets were added, since they were added uniformly throughout the battle space, more were added out of range than within range. Therefore, the collection statistics decreased as more targets were added. There were additional timing constraints for location and identification, which caused these probabilities to be less than the detection results.

The limiting effects of these timing constraints are illustrated in Figures 10.3 and 10.4. The probability of locating and identifying an NCS, shown in Figure 10.3, were about the same and varied from 0.2 at 100 target nets to about 0.1 at 300 target nets. P_{loc}, P_{det}, and P_{id} versus the number of target nets are shown in Figure 10.4. These values are somewhat lower than those for the NCS, consistent with the discussion above. P_{det} is not fully a factor of 1/2 that for the NCS. They vary from about 0.15 at 100 nets to about 0.07 or so at 300 nets. P_{det} is a little higher than P_{loc} and P_{id}, which are about the same. P_{loc} would be lower than P_{det} because of the data link cycle time of 8 seconds. Contention for DF assets caused some fixes to be lost even though the target was detected. P_{id} would be lower because not all of the targets that were detected were detected within the first 2 seconds. In fact, as discussed above, on average the earlier net members were rarely identified, while the latter net members, if the net was detected, were identified. The fact that P_{loc} and P_{id} turned out to be more or less the same is a coincidence. There is nothing in the basic parameters that would suggest this should happen.

10.4.2 Search Bandwidth

The wider the search bandwidth, the faster the search band was covered. The amount of external noise entering the system was obtained from the CCIR model described in

Chapter 3. In particular, in this case, a rural high noise level was assumed. These values are in decibels above *kTB*, where *B* is the system bandwidth, as described above. Therefore, the wider the search bandwidth, the more external noise entered the system. Thus, the tradeoff is speed versus signal level, both of which can limit the overall system performance at finding target signals.

In dense target environments a search bandwidth wider than the channel allocations allows for increased cochannel interference. In fact, cochannel interference is possible even if the target environment is not dense, but this cochannel, theoretically at least, is caused by friendly radios interfering with targets. With wide bandwidths, such interference can be caused by two friendly (or two or more) targets. There are 2,400 25 kHz channels between 30 and 90 MHz. With only a maximum of 300 target nets, the possibility of two adjacent channels being occupied is small (about 12%). However, when the bandwidth is 250 kHz, which is 10 channels wide, the possibility of cochannel interference in this case is a certainty (subject to the two nets operating at the same time). In any case, while cochannel interference is an operationally important consideration, the effects were not included in this simulation.

The systems started with no a priori information known about the threat environment. Therefore, at the beginning, all the searching was from a general search list that is of the "from f_1 to f_2 in ½ IFBW steps" variety. As time progressed and more targets were found, a more directed search was executed. In reality such a "cold start" would be highly unlikely because as much information about an adversary force as possible would be provided by higher echelon sources. This information would almost always include some data on net usage.

The results for $P_{\mathrm{id,NCS}}$ and $P_{\mathrm{loc,NCS}}$ versus the search bandwidth are shown in Figure 10.5. The variation as the bandwidth was increased is insignificant. This indicates that faster general search times do not increase the ability to find NCSs and it implies that the effects of increased noise were insignificant for this function as well. The same general characteristics for collection of target nodes in general are

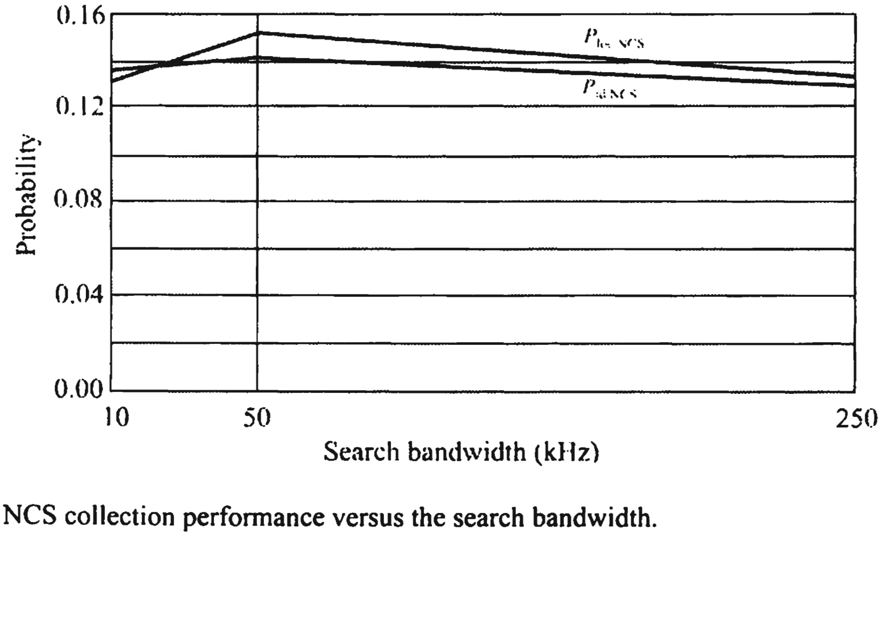


Figure 10.5 NCS collection performance versus the search bandwidth.

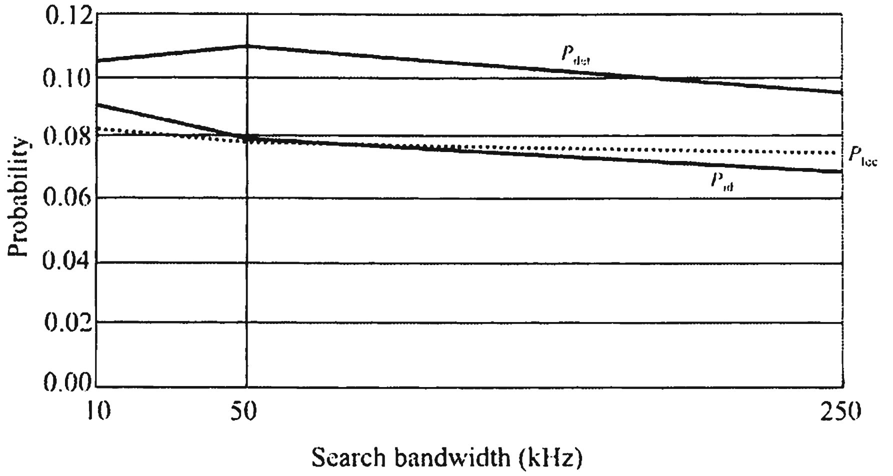


Figure 10.6 General collection performance versus the search bandwidth.

exhibited in Figure 10.6. The two opposing effects balanced each other out. There is a slight rise as the bandwidth increases from 10 to 50 kHz, implying that more targets were being found with a faster search rate, although the effects were slight. Yet the additional external noise is not that great. From 10 kHz to 50 kHz was an additional about 7 dB of noise. On the other hand, increasing from 50 kHz to 250 kHz, the collection performance dropped off, but again the differences are slight. This was an increase again of about 7 dB, but an increase of 14 dB from 10 to 250 kHz. The fact that these were offsetting effects can be seen from the results below on increasing the noise figure. In that case the search rate as well as the external noise were held constant because the bandwidth was constant. Increasing the noise figure did indeed decrease the collection performance. Hence, while the collection performance increased by increasing the search rate by widening the search bandwidth, it was more or less offset by the decrease in system sensitivity by admitting more noise.

P_{det}, P_{loc}, and P_{id} are plotted in Figure 10.6 versus the search bandwidth. The effects were the same as those indicated above for the statistics on collection of NCSs and for the same reasons. For all practical purposes these results show that it made no difference what the search bandwidth was. This is not really the case, however, for reasons that were not modeled here. The fact is that the range from the intercept systems was so small that using the narrowest search bandwidth is indicative of the limitation caused by an inadequate SNR. The wider the search bandwidth used, the greater the amount of noise present as discussed in Section 10.3. On the other hand, using a wider search bandwidth decreased the amount of time it took to search the RF spectrum for signals, thus increasing the probability of detecting a signal sooner. Clearly, it is better to use a narrow bandwidth and search more slowly, at least in a limited target environment as modeled here.

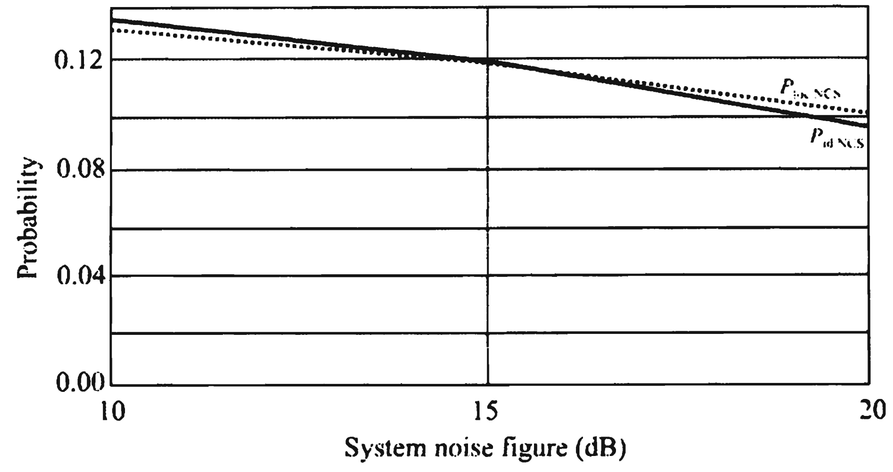


Figure 10.7 NCS collection performance versus the system noise figure.

10.4.3 Noise Figure

The system noise figure establishes the noise floor. This, in conjunction with the externally received noise, determines whether a received signal is detectable or not. If the received signal is below the minimum detectable signal level, then there is not enough energy present. Although low noise figures are typical for higher-frequency ranges, in the range of interest here, a 15 dB noise figure is considered typical, and 10 dB would be considered excellent. The noise figure is a design parameter that can be traded off with other design parameters. The probabilities of locating and identifying an NCS versus the system noise figure are shown in Figure 10.7. The effects of varying the system noise figure on collection in general are shown in Figure 10.8. These results show that the lower the noise figure, generally the better collection

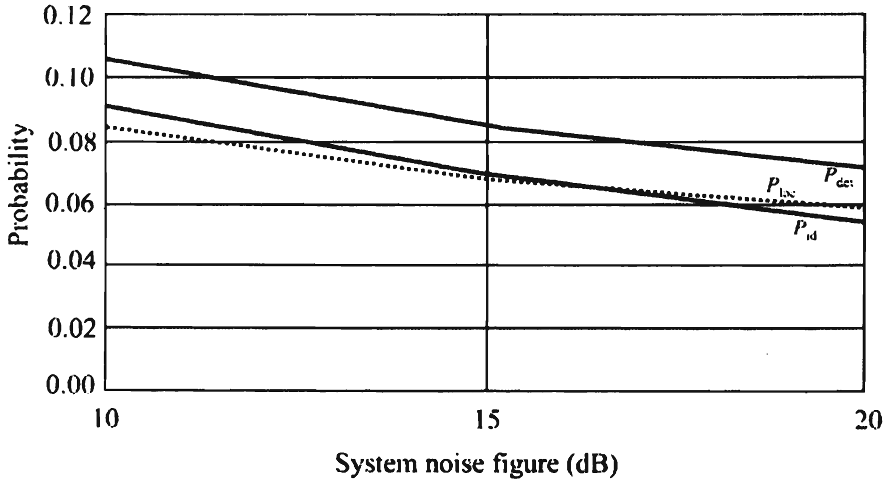


Figure 10.8 General collection performance versus the system noise figure.

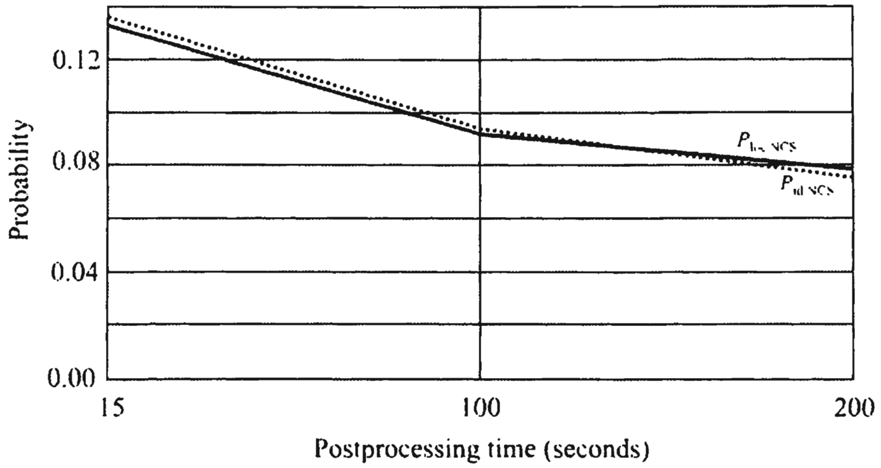


Figure 10.9 NCS collection performance versus the postprocessing time.

results ensued. A 10 dB noise figure generated about a 20% improvement in collection results over 15 dB.

The results shown in Figures 10.7 and 10.8 are somewhat surprising in one respect. In the frequency range considered, it is sometimes believed that external noise limits collection performance. Figures 10.7 and 10.8 show that this is not always the case for the noise figures considered, because lowering the noise figure without changing the external noise had a significant effect on collection performance.

As expected, as the noise figure was increased the probabilities decreased due to the reduced detection range caused by the additional noise. However, these results indicate that the noise figure needs to be considered in system design.

10.4.4 Postprocessing Time

The effects of varying the postprocessing time on NCS collection is shown in Figure 10.9 where $P_{\text{id,NCS}}$ and $P_{\text{loc,NCS}}$ are plotted. These two variables were about the same and varied between about 0.12 and 0.08. The gradual decrease in performance as the postprocessing time was increased was caused by the operators not being available to continue the general search. In fact, while postprocessing was occurring, there was no intercept going on—the signal had gone away. In early-entry operations, while in the GS mode during the first part of deployment, probably a message needs to be recorded on an intercept only if confirming the detection, location, or identification of an NCS. Otherwise, when very little is known about the target environment, reporting on every intercept is probably not necessary.

Collection performance against targets in general is plotted versus postprocessing times in Figure 10.10. The same general effects as far as collection of NCS targets are evident. About half the number of targets were collected at a 200 second

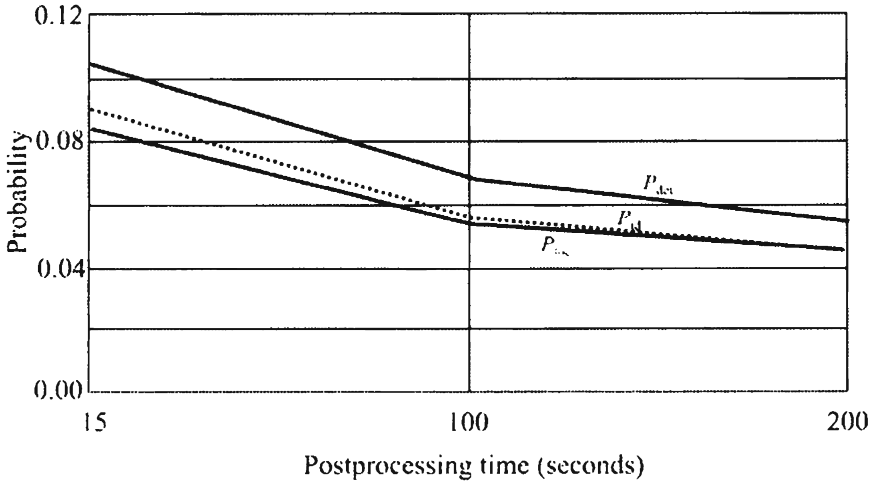


Figure 10.10 General collection performance versus the postprocessing time.

postprocessing time versus 15 seconds. This is also an indication that it was a target-rich environment for the systems; there were still targets to be collected within range of the EW systems.

If, when a net frequency was found, the collection for that net was assigned to an available operator somewhere in the collection ensemble, and, after that time, that operator only copied that frequency, it would be equivalent to using approximately three operators / 200 nets = 0.015 probability of identification and location of a net's control station in the 200 net case (ignoring for the moment hearability concerns). At a 15 second postprocessing time, the effects of searching and sharing collection assets increased the number of NCSs collected by a figure of about eight. At a 200 second postprocessing time this was reduced to a figure of about five. These figures indicate that the processing flow described earlier did indeed produce work enhancements.

10.4.5 Mission Duration

All else being equal (all targets are within range), as the duration of the mission is increased, more of the targets should be collected and that was indeed the case as shown in Figure 10.11. Approximately 75% more net control stations were collected at 3 hours versus 1 hour. Similar results were obtained for all of the targets as shown in Figure 10.12. These results indicate that within the first hour, neither all of the target net control stations nor all net members were detected, even though they were within hearability distance. In other words, as indicated above, it was a target-rich environment. There were still targets to collect and identify after the initial hour or two.

Recall that every so often, as determined by the revisit timer, the channels (frequencies) where targets had already been found were revisited. This tends to decrease the number of new target nets found, especially during the latter part of the

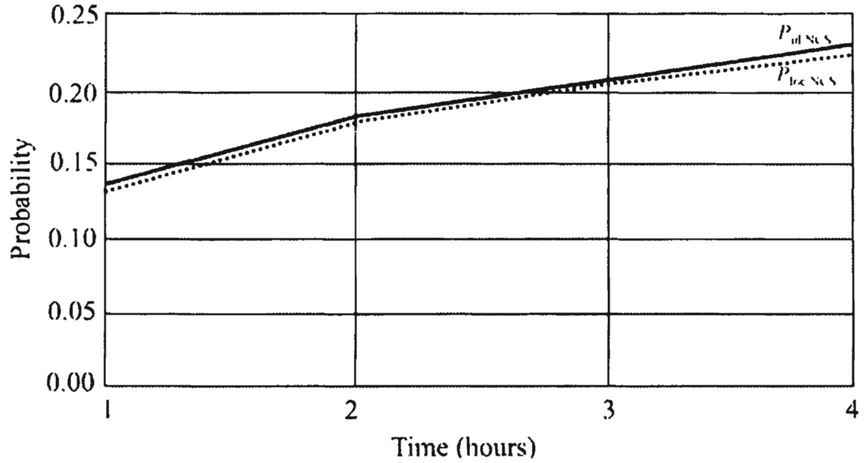


Figure 10.11 NCS collection performance versus the total mission time.

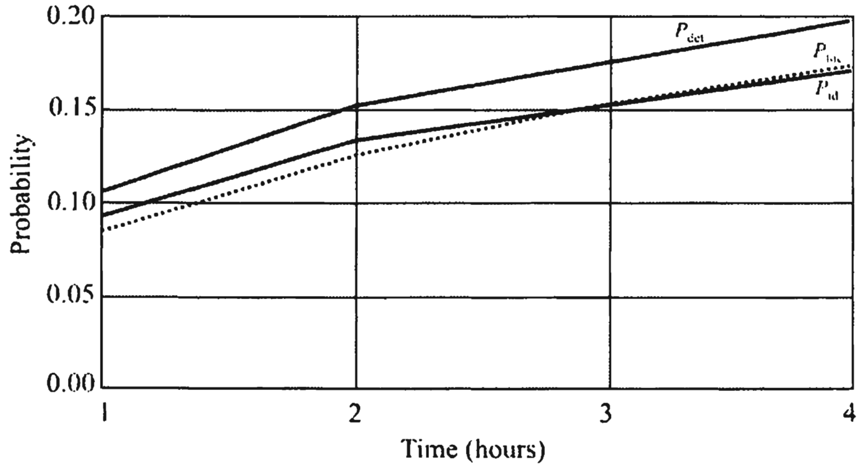


Figure 10.12 General collection performance versus the total mission time.

collection period, and therefore tends to decrease such probabilities computed here. In an operational scenario, as a net is determined to be of no interest it should be removed from the directed search list so it does not slow down the search process. Such frequencies go onto a "lock-out" list of frequencies. The revisit process, as determined by the revisit timer, is normally intended to provide intercept on frequencies known to contain activity. If it is not of interest it should be precluded.

This process is complicated by target frequency changes, however. It is common practice for communication networks to change their operating frequency on a regular basis to preclude or minimize targeted collection efforts such as those described here. It is necessary to recover the new frequencies once they have changed.

10.5 Concluding Remarks

One of the conclusions based on the results presented herein is that ground-based communication intercept systems should use a postprocessing time in the range of 20–100 seconds and no longer against limited target environments. The processing time after a signal has disappeared is comprised of two major components: (1) the requirement (or not) of preparing messages to be reported out of the collection system, and (2) whether the operator has to listen to a recording of the intercept in order to prepare the report.

The first component can be made smaller by not requiring extensive information to be reported from the collection system. This can be facilitated by having an off-line (to the collection mission), separate system with appropriate operators perform the reporting mission. This would necessitate the exchange of collected signals between the collection system and the reporting system, which could be accomplished in a variety of efficient ways. The second component can be made smaller by supplying the collection operator with more automated means to retrieve collected signals than analog tape recorders. Sophisticated digital storage means are available to allow instantaneous recall with random access to any speech segment. Providing the stored signal to a separate system from these digital recorders could easily be accomplished.

An alternative is not to require any reporting on signal content at all from such collection systems. In this case, the operator may simply edit fix data and nothing more, or raw fixes may be reported as is, with appropriate editing being performed elsewhere, if at all. In this case, postprocessing times would come very close to the optimum 20–100 second range indicated in the charts herein. Although this may be attractive because of its simplicity, it is highly questionable if such fix data alone would be of much value to anyone. In order to establish an adversary's intent and disposition, it is necessary to do more than just locate the targets on the surface of the Earth. Identification of a target, in general, requires that the contents of the transmission be understood. Other forms of automation support to collection operations, however, are also useful which collect external information about signals. Such techniques could include automatic translation and signal type identification.

Without a priori information about the target frequencies, it is necessary to search the frequency spectrum to find the active targets. No such knowledge was assumed for the simulation. Each operator was assigned an equal portion of the 30–90 MHz range and scanned that frequency band until targets were found.

Using a search bandwidth that is narrow is key to providing efficient collection. Herein, the effects on collection performance were essentially independent of the search bandwidth. Cochannel considerations as well as range considerations thus dominated the selection of search bandwidth, and these mutually required narrower bandwidths. In this case, searching more slowly is more important for noise elimination than searching rapidly by using a wide search bandwidth and allowing more noise.

The results of the simulation showed that external noise was not the limiting noise source in the frequency range considered (30–90 MHz). The noise figure of the receiving system should be as small as possible, and certainly should not be higher than 10 dB.

References

[1] TRADOC Pamphlet 525-5, "Force XXI Operations," Chapter 3, August 1994.

[2] FM 100-12, Passive Defense, Chapter 7.

[3] U. S. Army Posture Statement 2001, Chapter 3, The Army Vision and Force Modernization, accessed July 2001, http://www.army.mil.

[4] U.S. Department of Defense Joint Publication 3-0, Ch. IV, Joint Operations in War, June 1996.

[5] U.S. Army TRADOC Pamphlet 525-69, Concept for Information Operations, HQ, USA TRADOC, Ft. Monroe, VA, August 1995.

[6] U.S. Army Vision 2010, Dominant Maneuver.

[7] U.S. Army Field Manual 5-100, Chapter 12, Contingency Operations.

[8] U.S. Army TRADOC Pamphlet 525-75, Intelligence XXI, November 1996.

[9] *A Guide to the Use of the ITS Irregular Terrain Model in the Area Prediction Mode*, U.S. Department of Commerce, National Telecommunications and Information Administration, NTIA Report 82-100, April 1982.

[10] Adamy, D., *EW 101: A First Course in Electronic Warfare*, Norwood, MA: Artech House, 2001.

[11] Neri, F., *Introduction to Electronic Defense Systems*, Norwood, MA: Artech House, 1991.

[12] Torrerri, D. J., *Principles of Secure Communication Systems*, Norwood, MA: Artech House, 1992, Chapter 4.

Chapter 11

Detection and Geolocation of Frequency-Hopping Emitters

11.1 Introduction

Low probability of intercept communication techniques have been developed to try to thwart attempts at executing EW. One of those techniques is known as frequency hopping. In this technique, the frequency of the transmitter is changed often so that ES systems do not know where the frequency of transmission is at any point in time [1–3].

Described in this chapter is a technique for detecting and geolocating such frequency-hopping targets. The technique is first described analytically and then the results of a simulation of the technique for four different configurations of ES systems are presented.

11.2 Analysis

The technique consists of stepping a digitally controlled receiver from one channel to the next, dwelling on any one channel just long enough to measure the energy there. If the energy level exceeds an SNR threshold, then signal detection is declared. Once this occurs, other processing would be invoked, including geolocation of the target. This subsequent processing is discussed elsewhere herein.

The fundamental tradeoff in this technique is the scan rate of the ES receiver versus receiver bandwidth. As previously discussed, the wider the bandwidth, the lower the receive SNR for a constant signal level. Therefore, the wider the bandwidth, the fewer the number of targets that could be detected, and the lines of position are less accurate because of the lower SNR. However, more channels are covered in a given period of time, thus searching faster. On the other hand, the narrower the bandwidth, the slower the scan rate and therefore the fewer the number of channels covered during a given time period, taken as 1 ms and 10 ms for now. In this case, the SNRs are higher, however.

One of the more important questions surrounding a receiving system of the type described above is whether a scanning receiver will ever detect a signal and, if so, with what probability. By the nature of the processing, a single detection of a signal is adequate for locating the transmitter associated with it. Therefore, the probability of locating a target is equivalent to the probability of detecting it. This probability is calculated in this section.

A scanning superheterodyne receiver essentially forms what is called in communication theory a *partial-band filter-bank combiner* (PB-FBC) [4]. This is because at any given instant in time the receiver is dwelling at a single frequency with a given bandwidth. In the 10 ms that a 100 hps target dwells on a channel, the receiver covers one or more of these channels depending on the scan rate, as previously discussed. Therefore, the partial band corresponds to that portion of the 30–90 MHz frequency band over which the receiver is scanned in this time increment. The probability of a hit is given by

$$P_{\text{hit}} = \frac{N_{\text{s}}}{N_{\text{T}}} \tag{11.1}$$

where N_{T} is the total number of channels (25 kHz-wide channels in our case) in the total bandwidth and N_{s} represents the number of channels scanned in one time interval. Since the targets are assumed to be randomly hopping over the entire 60 MHz bandwidth and they do not use "hop sets," then $N_{\text{T}} = 2{,}400$. The values of N_{s} and the corresponding values for P_{hit} are given in Table 11.1.

The probability of getting n hits per transmitted message (for specificity a message corresponds to a 5 second transmission) is given by the binomial probability distribution, namely,

$$P_{\text{B}}(n) = \frac{N_{\text{h}}!}{n!(N_{\text{h}} - n)!} P_{\text{hit}}^{n} (1 - P_{\text{hit}})^{N_{\text{h}} - n} \tag{11.2}$$

which gives the probability of detecting exactly n target frequencies from a given transmitter. N_{h} is the number of hops in the message (in the case of interest here, N_{h} is equal to 500 since the messages are 5 seconds long and the targets are hopping at a 100 hps rate). This can be used to compute the probability of detecting at least one hop in a 5 second transmission for the various configurations because

Table 11.1 Partial-Band Bandwidths and Corresponding Values of P_{hit}

Dwell Time (ms)	Bandwidth (kHz)	N_s	P_{hit}
1	25	10	0.0042
1	100	40	0.0167
1	200	80	0.0333
10	25	1	0.00042
10	100	4	0.00167
10	200	8	0.00333

$$P(n \le N-1) = \sum_{n=0}^{N-1} P_B(n) \tag{11.3}$$

and

$$P(n > N) = 1 - P(n \le N-1) \tag{11.4}$$

These results are illustrated in Figures 11.1 and 11.2. The interpretation of these charts is as follows. The value on the curve at a value of the abscissa, say, n_1, represents the probability that n_1 (or more) hops will be detected in a 5 second transmission. The detection rules used herein only require the detection of one transmission to locate the targets; hence, the probability values beyond $n_1 = 1$ are of academic interest only. If a way could be found to correlate one fix with another, then multiple fixes on the same target could be used to improve the location accuracy, however. Clearly, when the dwell time is 1 ms, we are almost always assured of detecting at least 1 hop out of a 5 second transmission independent of the bandwidth used. For 10 ms, the probability of detection, and therefore location, is not as good for a single 5 second transmission but for the wider bandwidths is still respectable. It can be concluded that there is a good chance of detecting and, therefore, locating targets with a scanning receiving

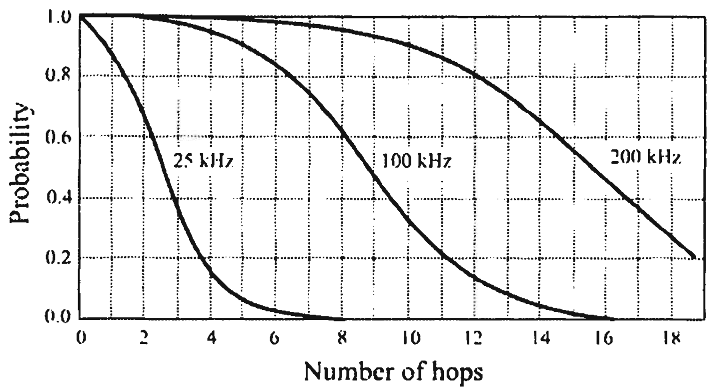


Figure 11.1 Probability of detection (location) of a target for a 1 ms dwell time.

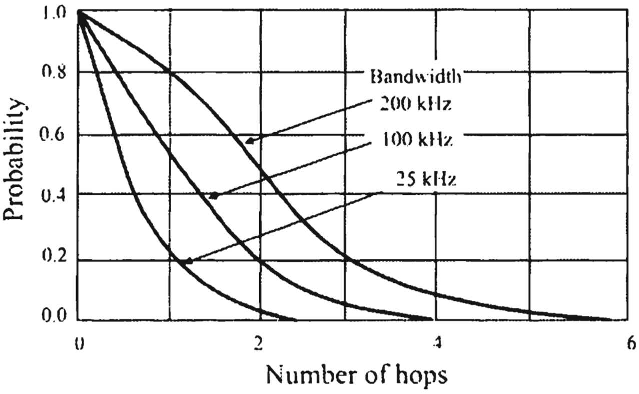


Figure 11.2 Probability of detection (location) for a 10 ms dwell time.

system that scans at a rate of between 1 ms and 10 ms per channel operating against 100 hps targets after only 5 seconds of collection. Furthermore, this can be accomplished using reasonable instantaneous scanning bandwidths to reduce problems caused by cochannel interference and sensitivity reductions due to noise.

11.2.1 Scanning Superheterodyne Receivers

This type of receiver is used for searching the spectrum, looking for energy in frequency channels. Thus, frequencies of signals of interest as well as an estimate of their amplitudes can be determined. Scanning receivers of this type have long been used to implement spectrum analyzers, a common piece of RF electronic test equipment. A block diagram of such a receiver is shown in Figure 11.3. Torrerri [5] provides a detailed analysis of this receiver. Lehtomaki et al. [6] discuss the performance of this receiver for FHSS signals.

The swept local oscillator causes signals within a frequency band to be mixed within the mixer. The preselector filters must also be tuned along with the local oscillator. A narrowband signal at the input will be mixed whenever the local oscillator tunes to the IF offset from the signal (recall that the mixer output is a

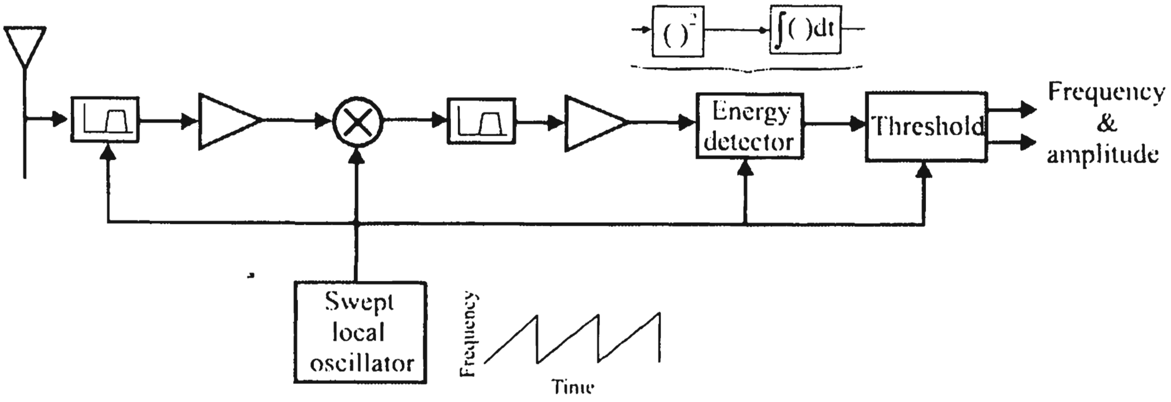


Figure 11.3 Block diagram of a scanning superheterodyne receiver.

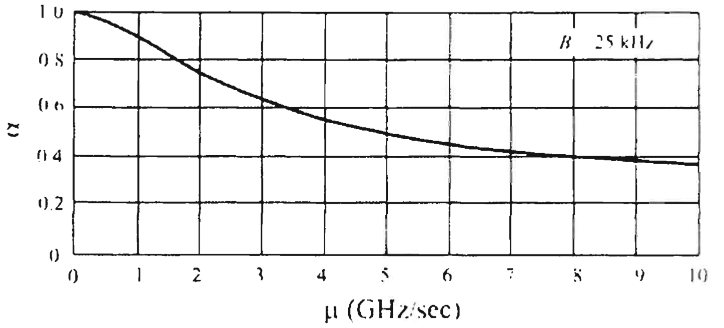


Figure 11.4 Example plot for the normalized peak value for a scanning superheterodyne receiver when B = 25 kHz.

constant frequency). At that point in time the AM detector will detect the peak amplitude of the signal and the peak detector will measure the amplitude. By measuring where the peak occurs in time, that time can be compared with where the scanning local oscillator is at that time so a frequency estimate can be computed.

As derived in Torrerri, if μ represents the scanning rate of the receiver in Hz per second and B represents the 3 dB bandwidth of the bandpass filter after the mixer, the normalized peak value α (normalized relative to the amplitude of the input signal) is given by

$$\alpha = \left(1 + 0.195\frac{\mu^2}{B^4}\right)^{-1/4} \tag{11.5}$$

and the frequency resolution is given by

$$\Delta = B\left(1 + 0.195\frac{\mu^2}{B^4}\right)^{1/2} \tag{11.6}$$

These expressions are plotted in Figures 11.4 and 11.5 when the bandpass bandwidth is 25 kHz. Thus the normalized amplitude peak value decreases as the scan rate increases, while the resolution increases in bandwidth (decreases in selectivity). In a dense RF environment a resolution bandwidth of less than 50 kHz or so is desirable in order to minimize adjacent channel interference. Thus, in the military VHF range where the signals typically have a bandwidth of 25 kHz, the scanning range must be kept at about 3 GHz per second or less in order to maintain the required resolution.

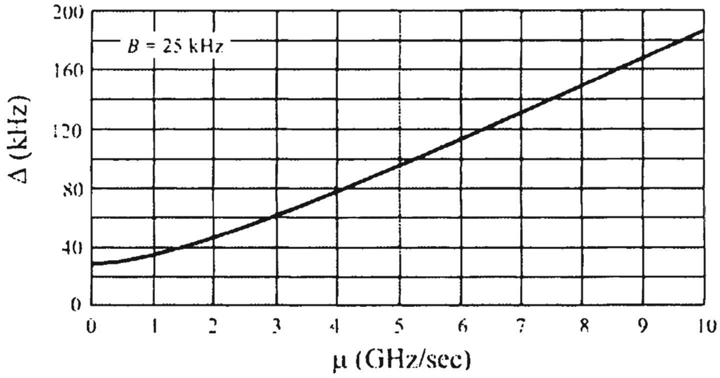


Figure 11.5 Scanning superheterodyne receiver frequency resolution when $B = 25$ kHz.

11.2.1.1 Scanning Superheterodyne Receiver Intercept Performance with Slow Frequency-Hopping Targets

The performance of a scanning superheterodyne receiver against slow frequency-hopping targets is analyzed in this section. The receiver is assumed to be a digitally tuned receiver that tunes a bandwidth B to a frequency and dwells there for a time given by τ_1. "Slow" refers to the hop dwell time compared to the receiver dwell time as opposed to the more usual definition of multiple chips per dwell. The receiver completes a scan across a predefined frequency band specified by f_1 to f_u, in time T_1, then resets and repeats this process. The target transmitter hops randomly anywhere in the total bandwidth as illustrated in Figure 11.6 and the frequency of the current hop dwell is denoted f_{tgt}.

The performance will be measured by calculating the *probability of intercept* (POI) of the target. Note that this is not the probability of detection normally associated with RF targets. That aspect is discussed in the next section. The poi is determined by whether a hop coincides with a receiver dwell (or vice versa), which is strictly a function of system and target timing. The latter is determined by the SNR at the receiver and the type of modulation of the target.

Actually Figure 11.6 is an easy way to visualize the operation but in the sequel the requirement to sequentially step through the spectrum is removed; the receiver can dwell in any part of the scanned spectrum—it need not follow sequentially from start to finish. This is a generalization of the situation depicted in Figure 11.6. It is not necessarily the best way to implement a scanning detector because conceivably the receiver may have to tune from one end of the spectrum being covered to the other. Most digital synthesizers would take longer to settle at the new frequency in this case as compared to tuning a signal bandwidth away. In any case, it is assumed that in T_1, the entire spectrum region of interest is searched with no segments repeated.

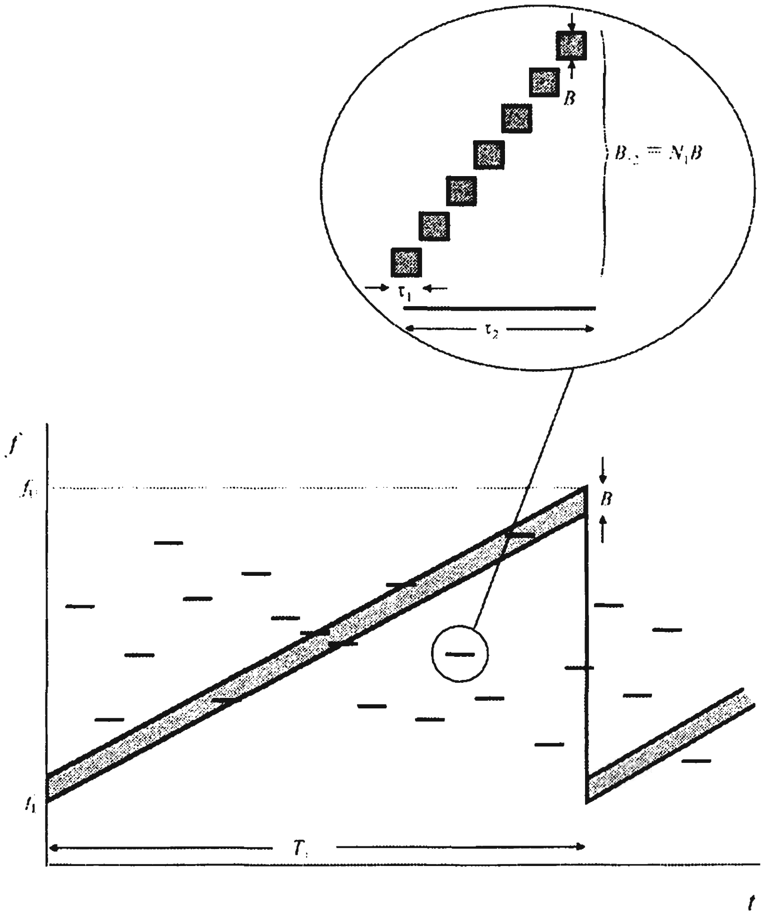


Figure 11.6 Frequency coverage over time with a scanning superhet receiver.

Although there are N_c = 2,400 channels 25 kHz wide in the low VHF spectrum from 30 MHz to 90 MHz, most SFH targets do not use all channels all the time. The group of channels actually used by a target is referred to as the *hop set*. The cardinality of the hop set is denoted by N_2 so that

$$N_2 < N_c \tag{11.7}$$

and the fractional size of the hop set is given by N_2/N_c.

Utilization of a hop set as a subset of N_c along with the pseudo-random usage of frequencies from the hop set leads to the target frequencies being nonperiodic. Hence, the approaches employed by several authors in the past [7–9], which assume that both the scanning receiver and target signal are periodic, are, in general, not applicable to the problem analyzed here.

Let τ_2 denote the target dwell time. For simplicity herein the time to hop to the next frequency is not considered. Not much is lost to generality with this assumption. The edge effects of not having the receiver dwell lined up with the beginning and end of a hop are not considered either. For the number of receiver dwells considered that should not make a large difference in performance. Let N_1 denote the number of bandwidth segments covered in τ_2 and let N_3 denote the number of bandwidth segments covered in T_1.

What is to be calculated is the probability that one or more of the N_1 receiver dwells in τ_2 includes the target frequency. Let this probability be denoted by P_1. This probability can be calculated by finding the probability that there are no receiver dwell segments that include the target frequency and subtracting that probability from one:

$$\begin{aligned} P_1 &= \Pr\{\text{One or more of } N_1 \text{ dwells includes } f_{\text{tgt}}\} \\ &= 1 - \Pr\{\text{None of the } N_1 \text{ dwells includes } f_{\text{tgt}}\} \end{aligned} \tag{11.8}$$

Let p denote the probability that f_{tgt} is included in at least one receiver dwell. Then

$$p = \Pr\{\text{Some receiver dwell includes } f_{\text{tgt}}\} \tag{11.9}$$

so that

$$p = \frac{N_1 B}{N_3 B} \tag{11.10}$$

Let q denote the probability that no receiver dwell includes f_{tgt}. Then

$$q = 1 - p \tag{11.11}$$

$$= 1 - \frac{N_1}{N_3} \tag{11.12}$$

Now, the probability that all N_1 dwells of the receiver are not included in the bandwidth $N_1 B$, implying that all receiver dwells are included somewhere else in the rest of the spectrum except $N_1 B$, is given by q^{N_1} so that, finally,

$$P_1 = 1 - q^{N_1} \tag{11.13}$$

Note that not all of these variables can be specified as some are dependent on others. For example, specifying τ_2 and N_1 determines τ_1 as illustrated in the following example.

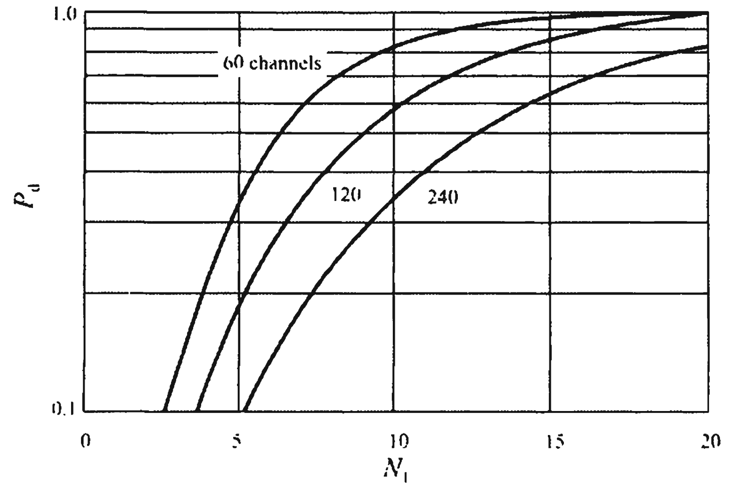


Figure 11.7 Detection performance of scanning superhet receiver.

An example will help clarify these notions. Suppose the region of concern is the low VHF spectrum from 30 MHz to 90 MHz mentioned above so that $N_c = 2{,}400$. Furthermore suppose the SFH target hops at a rate of 100 hps so that $\tau_2 = 10$ ms. Let $B = 1$ MHz so that $N_3 = 60$. The results for these parameters are shown in Figure 11.7. The results are also shown for when the spectrum is divided into 120, 500 kHz channels or 240, 250 kHz channels.

As shown, as the width of the receiver dwell decreases (N_3 increases), more dwells are required for a given POI. This is satisfying since as the number of dwells increases with a fixed dwell rate, each channel is visited less frequently.

It should also be noted that selecting system parameters as here can impact performance in other ways. For example, suppose the hop set consists of 250 frequencies with $B = 500$ kHz. Then to achieve $P_1 \geq 0.8$, $N_1 = 10$ or more. A target dwell is only 10 ms long, however, so to have the receiver step 10 times in 10 ms implies that it dwells for $\tau_1 = 1$ ms each time. This dwell time dictates a minimum frequency resolution of $1/10^{-3} = 1$ kHz. Signals located closer than this in frequency cannot be further resolved by subsequent processing. Of course, B could also be manipulated to achieve required performance levels.

In any event, further processing would be required, such as calculating an FFT of the signal over the bandwidth B, in order to achieve the necessary channel resolution to less than the channel width in the low VHF range (25 kHz presently). A 16k point FFT and a bandwidth of 1 MHz yield 62.5 Hz per point—more than adequate. Several such FFTs could be computed in 1 ms to take advantage of averaging to reduce noise.

11.2.1.2 Detection and Geolocation of Frequency Hopping Signals with a Sweeping Radiometer

The ability of a sweeping or scanning superheterodyne receiver to intercept, that is, overlap in time, a frequency-hopping target was described in the last section. That ability is strictly a function of the timing of the scanning receiver and the timing of the target. The ability to detect that target, on the other hand, is mostly a function of the modulation type and the SNR. This latter functionality is developed in this section. The receiver itself is assumed to implement a radiometer (energy detector). This section is based largely on the development in [6].

The measure of performance used here is the required SNR of the target signal at the receiver in order to achieve a specified probability of detection, given an acceptable false alarm rate. The results presented here are theoretical because they assume that there is only one target to be detected, which is not true in a tactical military scenario. They do, however, provide an upper bound on the performance of a radiometric receiver for detection of frequency hopping networks.

The search receiver scans or stares at a portion of the frequency spectrum to determine whether only noise or noise and signal is present at a frequency. For this discussion the scanning is accomplished by a superheterodyne receiver that is incorporated in a radiometer. The Neyman-Pearson detection criterion [10] is typically used to ascertain which of the two conditions exists at any given frequency. The optimum detector in the Neyman-Pearson sense is the one that gives the best probability of detection for a given probability of false alarm, which is also called *constant false alarm rate* (CFAR) detection.

A radiometer measures the energy or power of a signal. It integrates energy in a frequency band with bandwidth W_R and when the energy V integrated over the last T_R seconds exceeds a prescribed threshold. After that time the integrator is reset and the process repeats. A channelized radiometer integrates energy in many bands simultaneously by using multiple radiometers, usually tuned to adjacent portions of the spectrum. Detection is typically done by using a logical OR function [4, 11, 12], to combine different radiometer outputs and by summing the last M OR outputs. The final decision is made by comparing the accumulated sum value against a threshold k_M.

In this section we analyze a channelized radiometer that, as previously, steps the received frequency band K times within each hop, thus producing a scanning effect across the frequency spectrum. When $K = 1$ there is no scanning effect. Different methods to combine the channelized radiometer outputs are analyzed in [4]. These methods are logical OR–sum, sum–sum, and max–sum. Herein, only the first of these will be presented.

Signal Model and Receiver Structure

As above, the signal to be detected is frequency hopping signal with N_h non-overlapping hop channels. The per-hop bandwidth is W_h and the hop duration is T_h. Typically in the low-VHF frequency range W_h is 25 kHz and T_h is 10 ms. The targets are assumed to employ noncoherent BFSK modulation, which is typical for slow frequency-hopping networks. The noise is assumed to be white and Gaussian with one-sided power spectral density N_0. The intercept receiver has N radiometers, that form the channelized radiometer, each with bandwidth $W_R = W_h$. The receiver steps the received frequency band K times within each hop. For simplicity, we assume synchronization with the hop timing and that the frequency ranges of the individual radiometers match that of the hop channels [13, 14]. As discussed previously, in practice there is random splitting of the signal energy in both time and frequency.

The measured energies at the radiometer outputs are sampled after accumulating the energy (integrating or summing if discrete) for T_R seconds after which the center frequency is changed. This means that the total number of radiometer outputs within a hop duration is

$$N_{\text{eff}} = KN \tag{11.14}$$

These outputs are indexed so that in the first dwell the outputs have indices $1, 2, \ldots, N$; in the second dwell the outputs have indices $N + 1, N + 2, \ldots, 2N$, and so on (see Figure 11.8). This detection structure is analytically equivalent to N_{eff} radiometers with bandwidth W_H and integration time [13]

$$T_{\text{int}} = \frac{T_h}{K} \tag{11.15}$$

Figure 11.8 shows the intercept receiver detection structure for the case of $K = 3$ detection phases per hop duration. Sweeping makes the equivalent instantaneous bandwidth larger but the integration time per dwell is reduced. The total number of hops observed per decision is denoted by M. The probability of intercept per hop is

$$P_I = \frac{N_{\text{eff}}}{N_h} \leq 1 \tag{11.16}$$

because the hop frequency is assumed to be random. In the case shown in Figure 11.8,

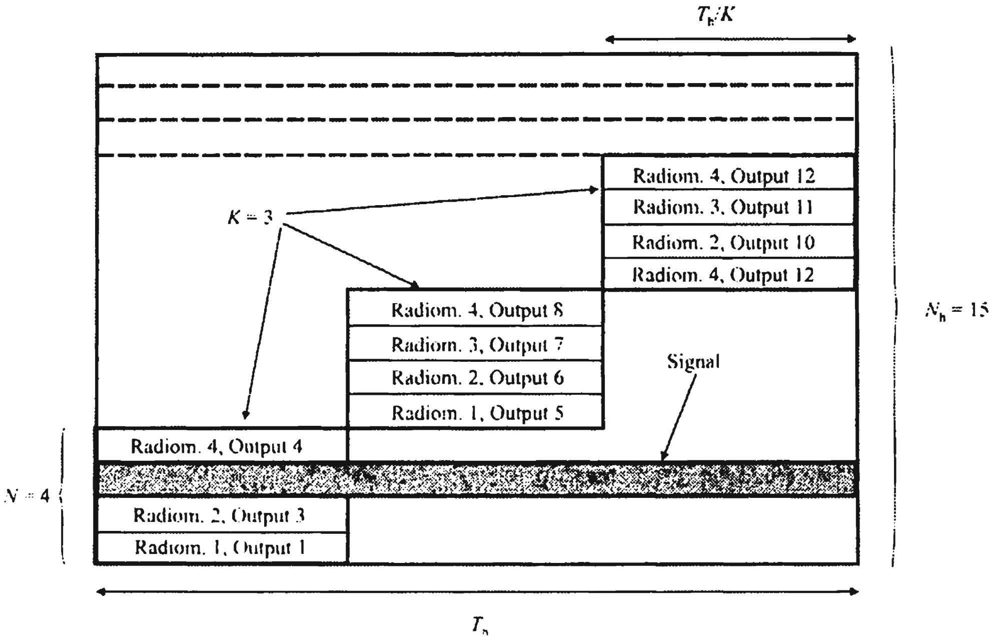


Figure 11.8 Detection structure in synchronous case, 15 possible hop channels, 4 radiometers, 3 detection phases, signal intercepted in first phase by radiometer 3, Neff = 11. (*Source:* [6]. © IEEE 2004. Reprinted with permission.)

$$P_1 = \frac{3 \times 4}{15} = 0.8 \tag{11.17}$$

The energy of the signal in the time-frequency region of the radiometer in which the signal occurs is given by

$$E_{ij} = \frac{E_\text{h}}{K} \tag{11.18}$$

where E_h is the received energy per hop, i is the hop index, $i \in \{1, 2, \cdots, M\}$, and j is the radiometer index, $j \in \{1, 2, \cdots, N_\text{eff}\}$. The time-frequency product of each individual radiometer is

$$T_\text{R} W_\text{R} = W_\text{h} \frac{T_\text{h}}{K} \tag{11.19}$$

and the combined instantaneous bandwidth of all radiometers is NW_h. In the following, it is assumed that $T_\text{R} W_\text{R}$ is an integer or that it is rounded to an integer.

Let

$$R_{ij} = \frac{2V_{ij}}{N_0} \tag{11.20}$$

where V_{ij} is the measured energy including the energy of the noise, denote a normalized radiometer output. In the noise-only case, the distribution function of R_{ij} can be approximated by the chi-square distribution with $2T_\text{R}W_\text{R}$ degrees of freedom. In the deterministic signal-and-noise case, the distribution can be approximated by the noncentral chi-square distribution with $2T_\text{R}W_\text{R}$ degrees of freedom and noncentrality parameter $\lambda = 2E_{ij}/N_0$. The *probability of false alarm* is the probability that the radiometer output exceeds a prescribed threshold, denoted by η, when only noise is present. Using the nomenclature of [13], individual radiometers have the probability of false alarm given by [13]

$$Q_\text{fa}(\eta) = \Pr\{R_{ij} > \eta | \text{signal absent}\} = \frac{1}{2T_\text{R}W_\text{R}\Gamma(T_\text{R}W_\text{R})} \int_\eta^\infty x^{T_\text{R}W_\text{R}-1} e^{-x/2} dx \tag{11.21}$$

where Γ is the gamma function. Using (11.21), the threshold for the required probability of false alarm can be determined (usually found by searching over values of η and calculate the Q's).

The *probability of detection* is the probability that the threshold is exceeded when both signal and noise are present which, assuming a deterministic signal is present, is given by [4]

$$Q_\text{d} = \Pr\{R_{ij} > \eta | \text{signal present}\} = Q_{T_\text{R}W_\text{R}}\left(\sqrt{2E_{ij}/N_0}, \sqrt{\eta}\right) \tag{11.22}$$

where Q_L is the generalized Marcum's Q function[1] with parameter $L = T_\text{R}W_\text{R}$. The SNR required for a specified probability of detection can be with [4]

$$\frac{E_{ij}}{N_0} \approx \frac{1}{2}\left\{\left[\sqrt{\eta - \frac{2T_\text{R}W_\text{R} - 1}{2}} - Q^{-1}(Q_\text{d})\right]^2 - \frac{2T_\text{R}W_\text{R} - 1}{2}\right\} \tag{11.23}$$

where $Q^{-1}(Q_\text{d})$ is the inverse of the detection probability (11.22).

[1] The generalized Marcum's Q function is defined by

$$Q_L(\alpha, \beta) = \frac{1}{\alpha^{L-1}} \int_\beta^\infty x^L e^{-\frac{x^2+\alpha^2}{2}} I_{L-1}(\alpha x) dx$$

where $I_a(x)$ is the modified Bessel function of the first kind and order a.

The FFT can be used to approximate an ideal channelized radiometer, with the FFT itself performing the function of a contiguous filter bank. An N-point FFT is computed every $1/W_h$ seconds so that the bin spacing corresponds to the FH channel separation [11]. There are $W_h(T_h /K)$ N-point FFTs per detection dwell. Squared magnitudes of FFT output bins are summed to get the channelized radiometer outputs. When only noise is present, the chi-square distribution is the exact result. Considerable leakage into adjacent output bins occurs if windowing is not used (recall that without windowing, the simple window function produces the sinc(x) response and the first sidelobe is only 13 dB down from the channel center). If windowing is used, correlation arises between FFT bins so that there is also correlation between the final radiometer outputs.

For the logical OR operation the probability of false alarm per hop is [11]

$$p_0 = 1-(1-Q_{\text{fa}})^{N_{\text{eff}}} \tag{11.24}$$

because a false alarm occurs if at least one radiometer output exceeds the threshold. At most one time-frequency cell (radiometer output) can have signal energy. The probability of this occurring is the probability of intercept, P_I. If the signal is not detected, the probability of detection is the false alarm probability. The total probability theorem can be applied to combine these two mutually exclusive events to get the probability of detection per hop as [11]

$$p_1 = p_I\left[1-(1-Q_{\text{d}})(1-Q_{\text{fa}})^{N_{\text{eff}}-1}\right]+(1-p_I)p_0 \tag{11.25}$$

Assuming that the hop channels are independent, the final probability of false alarm is [11],

$$P_{\text{fa}} = \sum_{i=\gamma_{\text{M}}}^{M}\binom{M}{i}p_0^i(1-p_0)^{M-i} \tag{11.26}$$

where γ_{M} is the final threshold that is used after summing the last M logical OR outputs. The final probability of detection is [11, 12],

$$P_{\text{d}} = \sum_{i=\gamma_{\text{M}}}^{M}\binom{M}{i}p_1^i(1-p_1)^{M-i} \tag{11.27}$$

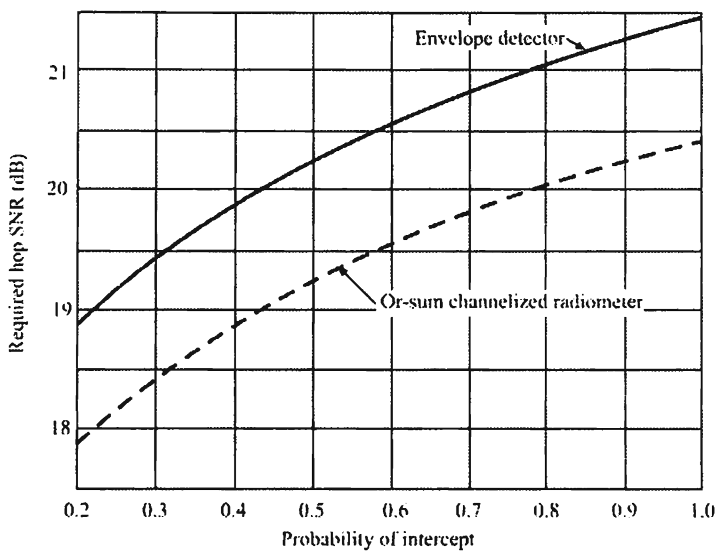


Figure 11.9 Required hop SNR (E_h/N_0), $W_h = 25$ kHz, $T_h = 0.01$, $N_h = 2{,}320$, $M = 300$, $P_d = 0.999$, $P_{fa} = 10^{-3}$. (*Source:* [6]. © IEEE 2004. Reprinted with permission.)

Example [9]. Assume that the signal to be detected is typical of the low VHF frequency range (30–88 MHz); namely, $W_h = 25$ kHz corresponding to $N_h = 2{,}320$ channels, and $T_h = 0.01$ seconds corresponding to a hop rate of 100 hps. Other parameters used here are $P_{fa} = 10^{-3}$ and $N = 464$ channels in the receiver. The probability of intercept for these parameters is $p_I = KN/N_h = 0.2K$. For example, when $K = 1$, so that no sweeping is performed, $p_I = 0.2$. We will assume that the spectrum is monitored for 3 seconds before a decision is made; thus the number of hops observed per decision is $M = 300$.

Figure 11.9 shows the required energy per hop to have $P_d = 0.999$ with envelope detector [12], and the logical OR sum-based channelized radiometer discussed here. Envelope detection is about 1 dB worse than the radiometer. We can see that when the number of hops observed is large, sweeping does not increase the probability of detection. This is in line with the result in [13], obtained for the envelope detector based system. However, if the frequency band of the transmitter is unknown, some type of frequency sweeping should be used. Otherwise, the target transmitter may not hop into the dwell width of the intercept receiver and the probability of intercept can be very small.

Note that this analysis is highly theoretical and useful for comparing detection techniques. It is not very useful for system design since it assumes there is only one target transmitter in the frequency spectrum under consideration, or that perfect isolation of the target of interest is possible over the observation interval (M hops) if there are multiple target transmitters. Methods and techniques for this isolation and association are not included in this analysis but are assumed.

11.3 Simulation

In order to ascertain the significance of the above analysis, a simulation was developed. The simulation assumed four different configurations of collection/geolocation systems. These four configurations are illustrated in Figure 11.10. For configuration 1, a low-flying airborne system operating with three ground standoff systems was simulated. The airborne system was at a standoff range of 20 km at an altitude of 5,000 feet. The ground systems were deployed in a lazy V configuration, equally separated vertically, and at standoff ranges of 5 km and 8 km as shown. In configuration 2, the airborne system was not deployed. The ground systems were deployed as in configuration 1. For the third configuration, two higher-flying airborne systems were simulated at a standoff range of 70 km and at an altitude of 20,000 feet. The last configuration is similar to configuration 1 except the ground systems were deployed in the midst of the *area of interest* (AOI), thus simulating a nonlinear battlespace.

The target networks in the simulation were based on the threat disposition shown in Figure 11.11, which was an area 50 km wide and 30 km deep. The nodes shown were configured into specific networks which comprise 19 specific

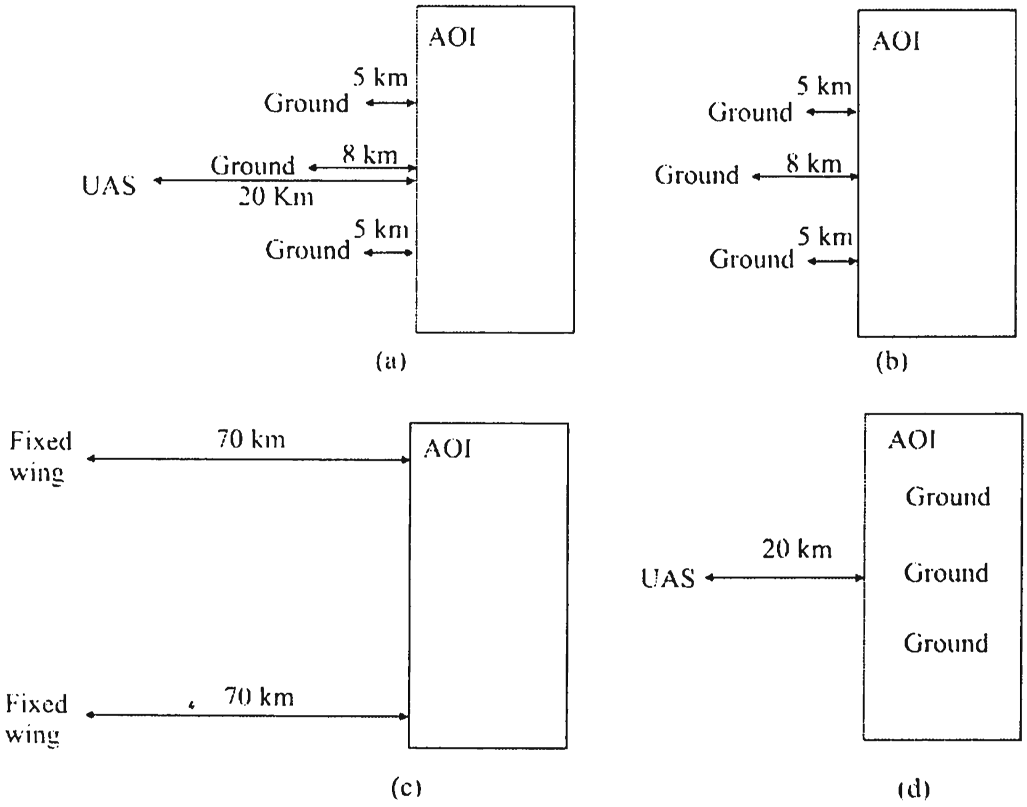


Figure 11.10 System configurations considered in the simulation: (a) configuration 1, three ground systems with one low-flying airborne sensor; (b) configuration 2, three standoff ground sensors; (c) configuration 3, two high-flying standoff sensors; and (d) configuration 4, three ground stand-in sensors with a remoted UAS.

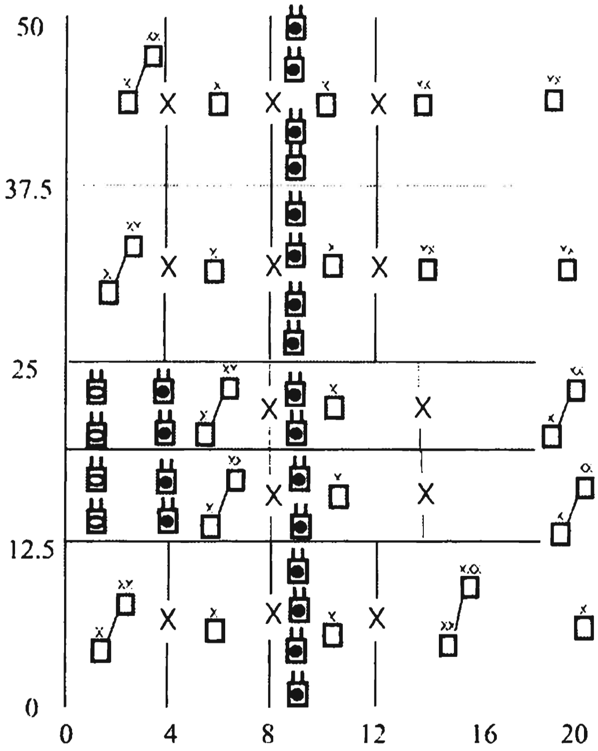


Figure 11.11 Target scenario for the simulation.

networks of radios. The networks are graphically portrayed in Appendix B. The target network communications are described in Section 10.2.

11.3.1 ES System Operation

The receiving systems consisted of stepping narrowband receivers, energy detectors connected to the output of these receivers, thresholding devices that measured the output of the energy detectors, and azimuth/LOP measuring devices. The processing time, that is, the time it took to do the stepping from one RF channel to the next, to perform the energy measurement and thresholding function and to measure the LOP was varied according to the configuration described below, between 1 ms and 10 ms. The basic operation is the same as described in Section 10.3, except the only function performed is direction finding.

The operational modes of configurations 1 and 4 are different from those in 2 and 3. The result of the propagation differences is that the ground stations do not receive much signal energy. The stations scanned the frequency spectrum in a synchronous way, that is they stepped to each channel at precisely the same time. In those cases where ground stations were included with an airborne element (configurations 1 and 4), each ground station, at each channel, computed an azimuth whether they detected the signal or not. This is feasible because, for correlation interferometry as well as some other types of DF techniques, usable LOPs can be obtained on signals that are weaker than can be automatically detected (the distinction here is signal detection versus computing an LOP) [15].

In configuration 2, even though an LOP was computed at the ground sensors regardless of the SNR present there, an adequate SNR was required at the DF net control station (the middle one) for signal detection, otherwise the systems would not know if there was a signal present or not. The randomly computed fixes that would result would be worthless information.

The accuracy of a computed LOP normally will vary proportionally to the inverse of the square root of the SNR. Depending on the particular DF approach used, a DF system may compute an LOP on signals that are too weak to be automatically detected. Some approaches even provide a capability at levels below 0 dB SNR. This is achieved by making the processing bandwidth very small, smaller than a VHF channel, thereby substantially decreasing the processing noise. For the purpose of the subsequent analysis, it was assumed that the Cramer-Rao accuracy bound was achieved in the LOP systems, an optimistic assumption. Note that wider receive bandwidths produce better accuracy; this, then, is another system tradeoff that must be considered, because the wider the system bandwidth the more noise enters, and therefore the lower the SNR.

The collection equipment on the low-flying system in configurations 1 and 4, which was flying at a considerably higher altitude than the altitude of the ground stations, operated in a similar fashion except that energy detection was first performed. Every station transmitted its time-frequency LOP information to the computer at a ground site, with the airborne sensor serving as a communication relay for the forward-deployed ground stations. The ground site only computed and displayed fixes at those frequencies at which the airborne sensor indicated there was energy present. For configuration 3, the operation was identical to that indicated above for the low-flying system with energy detection being performed at each site before an LOP was computed. In all of these configurations, each receive site computed LOPs and sent this information to a central location where a fix was computed.

The digital receivers were stepped with instantaneous bandwidths of 25, 100, and 200 kHz. The analysis provided above calculated the probability of detecting a hopping target with such a technique. The scan rate was different for configurations 1 and 4 then 2 and 3. This is because large quantities of data must be transferred when azimuths are reported at every scanned channel and the data link would not support the same scan rates. For configurations 1 and 4, a 10 ms dwell time per channel was used, whereas in configurations 2 and 3, the systems dwelled for 1 ms per channel. To put these numbers in perspective a receiver with a 10 ms dwell time could scan one channel in 10 ms (the time a frequency hopper stayed at a single frequency) while with a 1 ms dwell time, 10 channels could be measured. In the latter case, when the bandwidth was 25 kHz, a total of 250 kHz could be covered; when the bandwidth was 100 kHz, 1 MHz could be covered; and when the bandwidth was 200 kHz, 2 MHz could be covered.

A target location was obtained on each cell according to the rules explained above, as long as adequate data link capacity was present. The rate of the data link that interconnected the systems together limited the number of LOPs that could be transferred among systems. Since all the systems were time synchronized and covered the same instantaneous frequency band at the same time, a target detected at one system at a frequency would also be detected at the same time at all collection systems, subject to signal level constraints (the signals must have been of an adequate level as well as an adequate SNR to be detected). If too many signals were detected at a time for the data link to transfer the LOPs to the analysis station, the remaining signals were considered not detected for the purpose of computing locations.

The data link data rate used herein was 200 kbps for configurations 2 and 3 and was 400 kbps for configurations 1 and 4. This is not the sustained rate, because all stations shared this capacity and in some cases relaying was necessary. The maximum number of LOPs that could be transferred over this data link was data rate/100 × 1/hop rate assuming that 100 bits per LOP record would be required. Since the hop rate was always 100 hps, then the maximum number of fixes that could be computed was 20 during any individual dwell of the receiver for a 200 kbps rate and 40 for a 400 kbps rate. The data link was assumed to be using a frequency of 1.7 GHz, and its propagation characteristics were included in the model.The collection system and target transmit antenna heights are among the variables that determine the amount of power received at the receive sites. The receiver antenna heights of the airborne systems were the altitude of the aircraft indicated previously. For the ground systems, in configuration 1 the receiver antenna height was 10m, while for configuration 4 the receiver antenna heights were 5m.

A fix may not have been obtained on a hop for a variety of reasons, probably the most important being that the channel was not covered by the receive systems at that dwell. Other reasons why an individual hop may not have collected were: the SNR was too low [an SNR of 15 dB was necessary which is approximately that required for an ideal energy detector to obtain a probability of detection, P_d, of 0.9 with a false alarm probability, P_{fa}, of 10^{-6}, and varying amounts of background noise as well as the receive system noise figure (10 dB) were added to the signals], or the signal level was too low (the receiver dynamic range was a simulation variable and close interferers combined with this dynamic range set the minimum signal levels that could be detected). In all cases a dynamic range of 72 dB for the receive systems was used. In the case of the airborne systems, the closest interferer was assumed to be a distance away equal to the altitude of the aircraft. For the standoff ground stations, the closest interferer was assumed to be 0.5 km away and for the cross-FLOT ground stations, the closest interferer was assumed to be 15 km away (close transmitters are not interferers but are targets, although close targets could still desensitize the receiver).

The receiver antenna gain used in the simulation is shown in Figure 11.12. An antenna with this gain characteristic is typical for the frequency range of interest herein. This antenna performance could be attained with a half-wavelength dipole tuned to a frequency of about 60 MHz, for example. This would be typical of a ground configuration but would probably be difficult to achieve with a small airborne platform, such as a UAS.

11.3.2 Results and Analysis

The results of the simulation are presented in this section. Two figures of merit were calculated for each configuration: the probability of location, P_{loc}, and the average fix miss distance. Both of these parameters are dependent on system timing parameters, which are the primary variables of interest in this analysis.

11.3.2.1 Probability of Location

Figure 11.13 shows P_{loc} for configuration 1 (low-flying airborne and three ground standoff systems) plotted versus bandwidth. The upper curve shows P_{loc} for the net control stations while the lower curve corresponds to all of the targets. In the simulation the net control stations broadcast much more frequently than the rest of the net members, which accounts for the difference (see Figure 10.1). Palpably, very good location performance is obtained in all cases except perhaps for using a 25 kHz bandwidth.

Figure 11.14 shows the same data for configuration 2 (ground standoff systems only). In this case the performance was not very good. This is because of

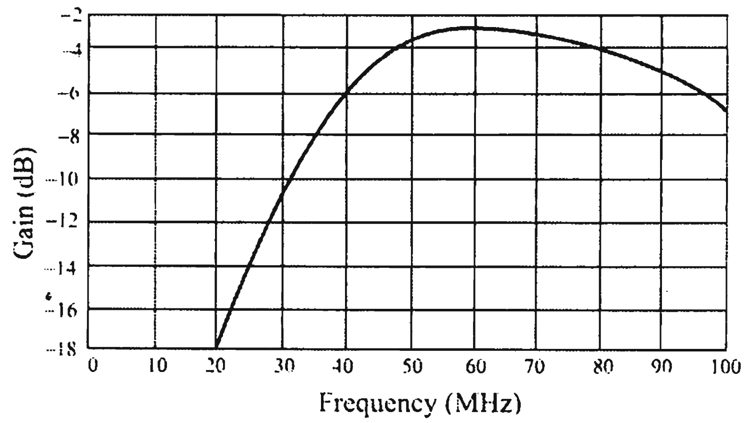


Figure 11.12 Receiver antenna gain used in the simulation.

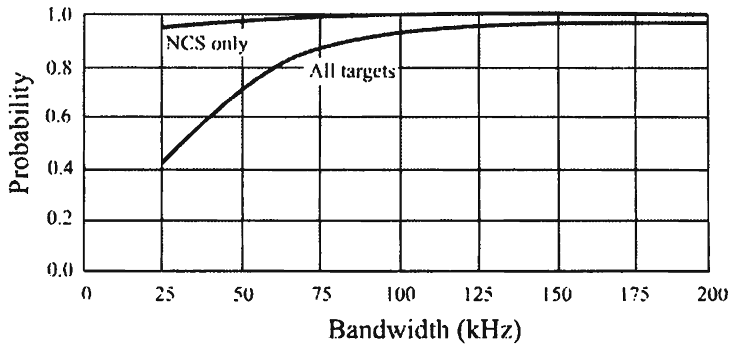


Figure 11.13 Probability of location for configuration 1, low-flying airborne with three ground standoff platforms.

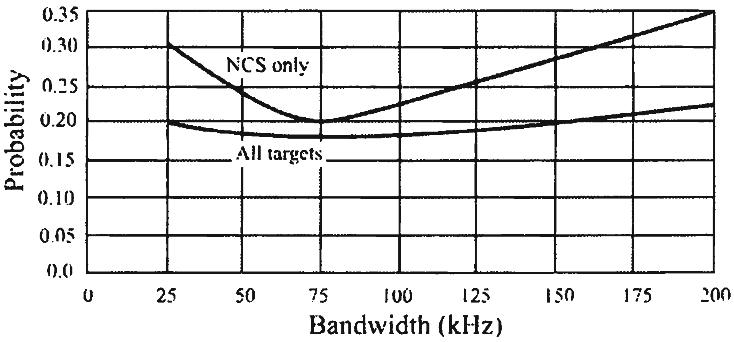


Figure 11.14 Probability of location for configuration 2, three ground standoff sites.

the low SNRs available at the ground stations for many (probably most) of the targets. The ground systems computed an LOP regardless of the SNR at the sensor; at the PF net control station the signal had to be of adequate SNR to be detected, thus the low probabilities of location (detection). One would expect these curves to be monotonically increasing functions of bandwidth, while Figure 11.14 shows a dip as the bandwidth is increased. The reason for this is the tradeoff between search speed; the wider the bandwidth, the faster the search and the higher the noise level thus decreasing the SNR.

For configuration 3 (two higher-flying aircraft), the results are shown in Figure 11.15. Here again, P_{loc} was very good; every NCS and almost every target was located regardless of the bandwidth. This was due to the relatively high SNRs that such a configuration would receive even though it was located 70 km to the rear of the FLOT. One would also expect that the LOP accuracy was good in this configuration. The fact that reasonable performance was achieved even at a bandwidth of 25 kHz is important because this configuration would experience the highest probability of cochannel interference.

Figure 11.16 shows the performance for configuration 4 (low-flying aircraft with three ground systems deployed 15 km across the FLOT). As with configuration 1 in this case, the ground sites did not have to detect energy first, but computed an azimuth on every channel. The low-flying aircraft performed the energy detection function. The performance was similar to configuration 1, which was expected because the detection functions of the systems are the same.

11.3.2.2 Miss Distance

We would expect that at least for configurations 1 and 4 the accuracy of computing fixes would be significantly degraded due to the fact that the ground sites could not receive an adequate SNR in many cases. A comparison of the average miss distances computed during the simulation is shown in Figure 11.17. The data in this chart represents the radius of a circle within which one-half of the computed fixes would lie and one-half would be outside. Configurations 1 and 4 produced the least accurate fixes, while configurations 2 and 3 produced the best. The decreasing LOP accuracy effects of the bandwidth (and therefore the SNR) are most evident for configuration 1 as the miss distance about doubles as the bandwidth increases from 25 kHz to 200 kHz. In the other cases apparently the SNRs were adequate most of the time to not significantly degrade the LOP performance.

As discussed in Section 11.3.2.1, one would expect these curves to be monotonically increasing with bandwidth because the LOP accuracy is inversely related to the square root of the SNR. The average miss distance for configuration 4 does not follow this pattern.

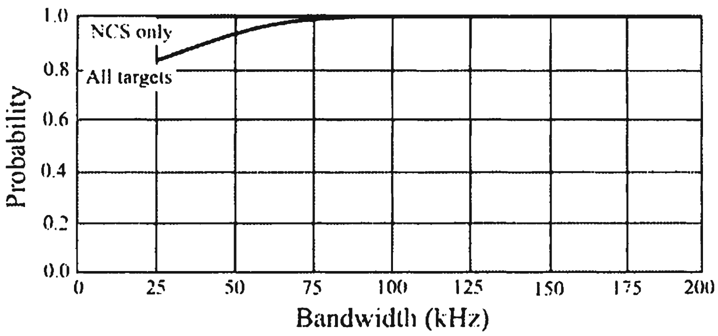


Figure 11.15 Probability of location for configuration 3, two fixed-wing aircraft.

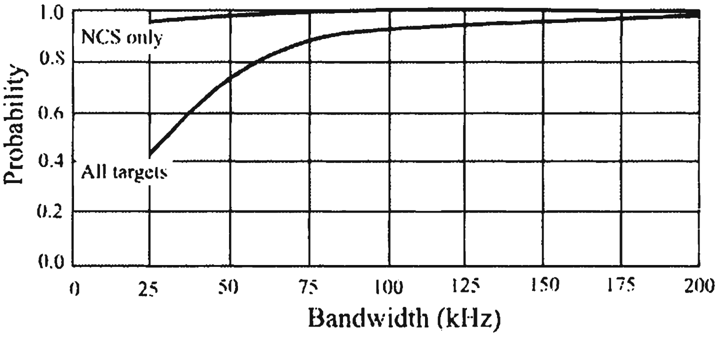


Figure 11.16 Probability of location for configuration 4, airborne sensor with three cross-FLOT sensors.

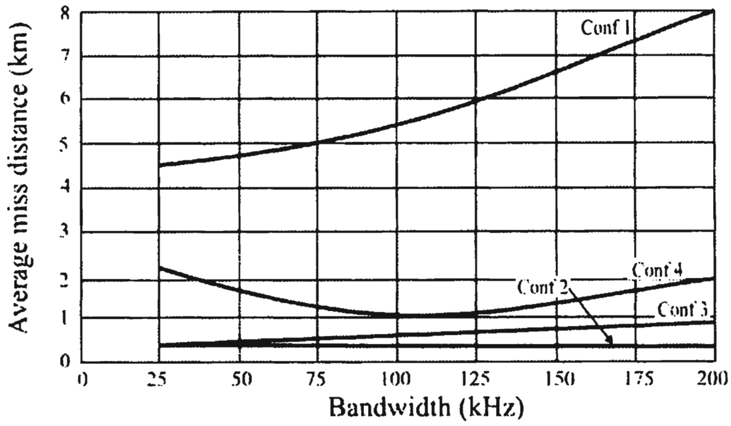


Figure 11.17 Average miss distance.

11.3.3 Discussion

One should not conclude which configuration performed the best overall based on one of these charts alone; they must be considered together. Configuration 2 produced the highest accuracy but missed many of the targets, as shown in Figure 11.17. The best overall results were obtained with the higher-flying aircraft configuration while the second best were obtained with the low-flying aircraft working with cross-FLOT ground systems.

One of the advantages of a DF system of the kind described herein is that its performance is not limited by the size of the target environment, except for when cochannel interference occurs. Collection systems that try to track LPI targets exhibit a characteristic that there is a point where the processing equipment cannot keep up with the load and they become saturated. Since LOP gates are often used as a sort parameter, basic system performance such as LOP accuracy also limits the sorting and therefore tracking performance of collection systems as well. Since the equipment analyzed here is not trying to track, it simply computes an azimuth at a frequency channel and reports the result; no such limitations exist.

While a 1 km miss distance when locating targets is probably not adequate for targeting purposes, it is adequate for target and situation development. Since the total target area covered was 20 km × 50 km, this corresponds to a total area of 1,000 km^2, the area covered by a 1 km circle is approximately 3 km^2. Therefore, it is possible with a system such as that described herein to localize targets to a granularity of 0.3% of the AOI at least half the time.

As expected, the case where the ground standoff systems computed an LOP on every channel ·regardless of the level of energy present produced the least accurate fixes. It is also the case, as expected, that when there was only ground standoff sensor systems they produced the fewest number of fixes because the distant targets could not be heard. These results indicate that whenever possible it is always desirable to have an elevated sensor component as part of the sensor array.

11.4 Concluding Remarks

11.4.1 Scanning Superheterodyne Receiver

The intercept performance of a scanning superheterodyne (narrowband) receiver was presented, with particular emphasis on the detection performance against slow frequency-hopping target networks. It was shown that detection (and therefore, with the assumptions in this chapter, target location) with a scanning receiver is entirely possible, and yields results at reasonable signal levels.

It was also shown that when the observation is over a large number of hops scanning the receiver produced diminishing returns. It was noted, however, that if the frequency band of the FH target networks was not a priori known, then some amount of scanning is required or else the intercept probability could be quite low, even zero.

11.4.2 Simulation

The results contained herein indicate that (1) theoretically there is a very high probability that 100 hps frequency-hopping targets can be detected by a scanning receiving system and (2) there are reasonable tactical configurations of receiving systems that can locate these targets. The accuracy of these locations varies, but one would expect that on the order of 1 km accuracy is possible.

The scan rate (bandwidth) versus SNR results can be summarized as follows. Any configuration with an aircraft involved produces $P_{\text{loc}} \rightarrow 1$ for some bandwidth, indicating that an adequ*a*te SNR was obtained for the scan rates used. For the ground-only case, SNR limits performance in most cases for any bandwidth.

References

[1] Simon, M., et al., *Spread Spectrum Communications Handbook*, New York: McGraw-Hill, 1994.

[2] Torrieri, D. J., *Principles of Secure Communication Systems*, 2nd Ed., Norwood, MA: Artech House, 1992, Chapter 1.

[3] Petersen, R. L., R. D. Ziemer, and D. E. Borth, *Introduction to Spread Spectrum Communications*, Upper Saddle River, NJ: Prentice Hall, 1995.

[4] Dillard, R. A., and G. M. Dillard, *Detectability of Spread Spectrum Signals*, Norwood, MA: Artech House, 1989.

[5] Torrieri, D. J., *Principles of Secure Communication Systems*, 2nd Ed., Norwood, MA: Artech House, 1992, pp. 323–328.

[6] Lehtomaki, J. J., M. Juntti, and H. Saarusarri, "Detection of Frequency Hopping Signals with a Sweeping Channelized Radiometer," *Conference Record of the IEEE Thirty-Eighth Conference on Signals, Systems, and Computers* 2004, Vol. 2, November 7–10 2004, pp. 2178–2182.

[7] Clarkson, I. V. L., J. E. Perkins, and M. Y. Mareels, "Number/Theoretic Solutions to Intercept Time Problems," *IEEE Transactions on Information Theory*, Vol. 42, No. 3, May 1996, pp. 959–971.

[8] Kelly, S. W., G. F. Noone, and J. K. Perkins, "Synchronization Effects on Probability of Pulse Train Interception," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 32, No. 1, January 1996, pp. 213–220.

[9] Clarkson, I. V. L., "The Farey Series in Synchronization and Intercept-Time Analysis for Electronic Support," *Transactions AOC* 2004, October 2004, pp. 7–28.

[10] Poisel, R. A., *Foundations of Communications Electronic Warfare*, Norwood, MA: Artech, 2008, Section 3.6.2.

[11] Miller, L. E., J. S. Lee, and D. J. Torrieri, "Frequency-Hopping Signal Detection Using Partial Band Coverage," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 29, No. 2, pp. 540–553, Apr. 1993.

[12] Dillard, R. A., "Detectability of Spread-Spectrum Signals," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 15, No. 4, July 1979, pp. 526–537.

[13] Levitt, B. K., M. K. Simon, A. Polydoros, and U. Cheng, "Partial-Band Detection of Frequency-Hopped Signals," *Proceedings IEEE Globecom* 1993, Vol. 4, Houston, USA, Nov./Dec. 1993, pp. 70–76.

[14] Levitt, B. K., U. Cheng, A. Polydoros, and M. K. Simon, "Optimum Detection of Slow Frequency-Hopped Signals," *IEEE Transactions on Communication*, Vol. 42, No. 234, pp. 1990–2000, Feb./Mar./Apr. 1994.

[15] Jenkins, H. H., *Small-Aperture Radio Direction-Finding*, Norwood, MA: Artech House, 1991.

Chapter 12

Signal Detection Range

12.1 Introduction

This chapter examines the distances over which communication ES systems can detect target signals. This range is dependent on several factors, including ES system sensitivity, antenna gains, power levels, and frequency. Determining the nominal detection ranges of typical systems against typical targets provides some insight to the limiting factors involved.

The detection range is also strongly dependent on how high the antennas involved are elevated. Herein, only ground-based targets will be addressed, so the only possible elevated antenna is on the receiving system. Four configurations are examined—the first is a high-flying aircraft. The second configuration is a low-flying aircraft, consistent with the flight characteristics of a UAS. The last two configurations are ground-based. For the first of these, the antenna is mounted on top of a relatively short mast—say, 10m or so in height. This configuration would be consistent with a stationary deployment. The second ground-based configuration has a short antenna mast such as that which would be used for *on the move* (OTM) operations [1–8].

The advantages of the utilization of UASs in many circumstances are obvious. They can overfly unfriendly areas with minimal risk of loss of life. They can be flown for long periods of time without the human-needs problems of manned aircraft. They can carry a varied mix of sensors, and, indeed, the sensors can be changed relatively easily.

The advantage of ground-based systems is their all-weather capability. Their most significant shortcoming is the limited detection range they provide. Even when placed on the highest hills available in the area, their range is exceptionally limited compared to their airborne counterparts. The rare exception would place a ground-based system on the edge of an available cliff with low-lying plains over which to watch—not too likely an occurrence, and even if it did occur, in conflict situations an adversary would not permit such systems to remain in operation very long.

Up until recently, the most significant advantage of high-flying, large aircraft was their weight-carrying capability. With the advent of large UASs, however, this advantage is disappearing. In some cases the ES system operators fly on the aircraft and in some cases they do not. In the latter case, the aircraft mission payload must be electronically linked somehow to wherever the operators are located.

Relatively high-flying fixed wing aircraft have been used for ES missions in the military services for several years. The higher above the surface of the Earth an ES system flies, the more signals can be seen, all else being equal. There are some limitations to using this configuration for ES, however. Its primary advantage can sometimes also be its biggest nemesis. Flying higher, more of the friendly transmitters can be seen, as well as targets. This creates cochannel interference. Depending on the type of signals involved and the geographical relationships among the emitters, this cochannel interference can sometimes be suppressed.

When possible, the high-flying systems are usually flown in a standoff posture—out of harm's way. In that case, the systems must look through many blue emitters before seeing the target array. To further exacerbate the problem, the signals from these blue emitters are frequently stronger than those from the red targets. In nonlinear situations there are also many blue signals mixed in with the targets. In mountainous terrain, the terrain can block communication signals. This is true for the intended receiver as well as an ES system. This can create coverage problems for standoff systems. In order to cover valleys in such terrain, over-flight is often the only way. Terrain effects can be helpful or harmful, depending on where the targets are located. Herein, terrain is not taken into account; smooth Earth is assumed.

In the future, low-power PCSs based on commercial technology are expected to proliferate in military and paramilitary forces worldwide. Such systems are characterized by very low-powered handsets. In addition, these systems intentionally reuse the frequency spectrum. Currently in the United States, every fourth cell in any straight line utilizes the exact same set of frequencies as illustrated in Figure 12.1. In Figure 12.1 the three cells marked with an X all utilize the same frequency set. The hexagonal cells, which of course in real systems only approximate hexagonal, arranged in this pattern, are maximally separated in distance from the next one that reuses the same frequencies. The seven cells surrounded in bold lines make up a "cluster" of cells. All the frequencies in a region are used in each cluster. If an ES system sees two or more of these cells that reuse frequencies, the same cochannel problem mentioned above occurs, degrading both the ability to intercept as well as the PF of a target.

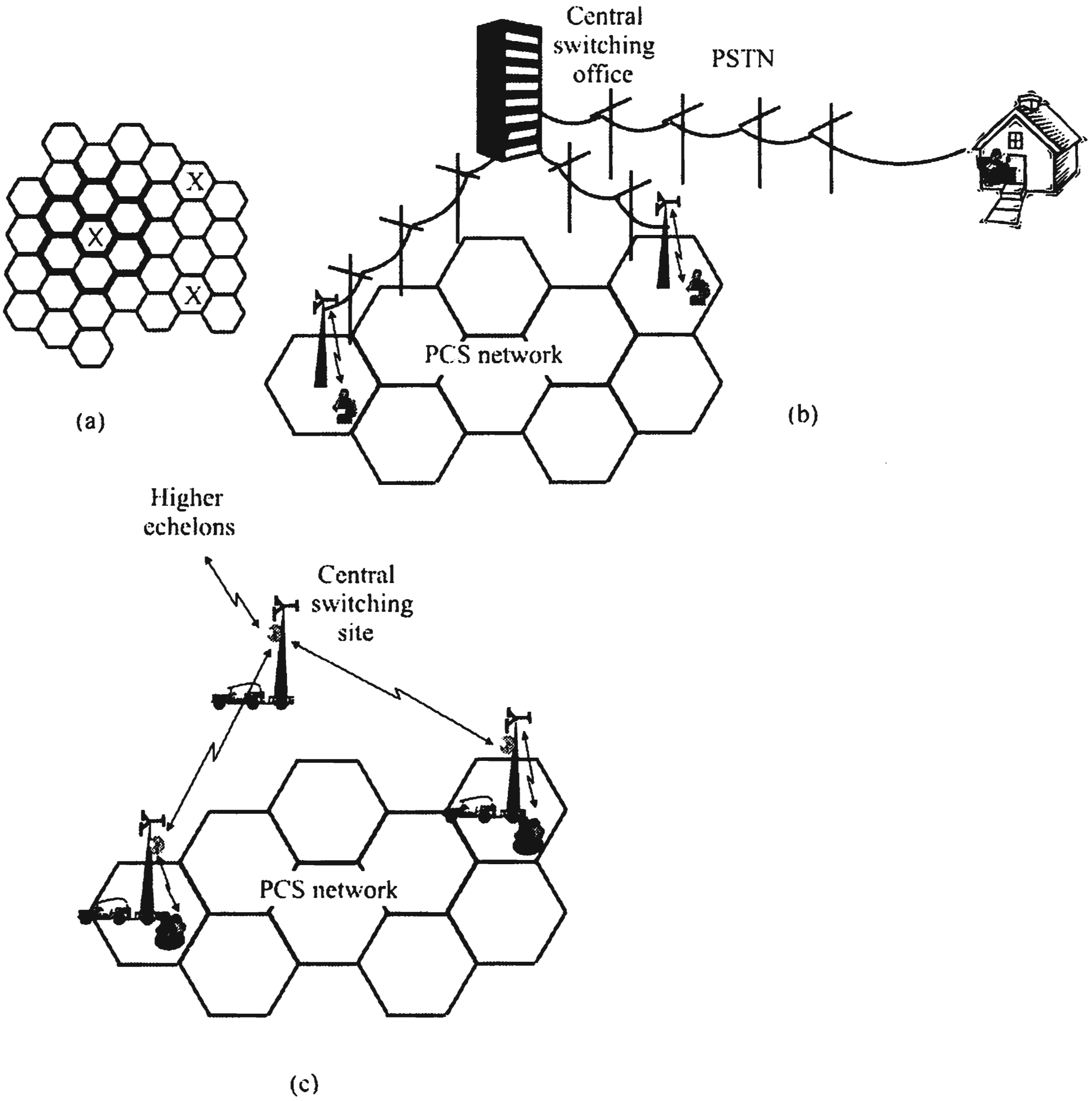


Figure 12.1 PCS systems—regional configuration of several PCS cells make up one system: (a) Those marked with X all reuse the same frequencies. The cells in groups of seven use all the frequencies in a region. (b) The physical makeup of a PCS system connects mobile users to other mobile users or mobile users to the PSTN. (c) Military and paramilitary PCS systems are similar to their commercial counterparts.

Therefore, low-flying systems or ground-based systems are required in such instances [4, 9]. PCS base stations will have substantially more power available than the handsets. For commercial communication systems the base stations are fixed sites. When military they will be mobile but not normally moving when in operation. Thus, they can facilitate operation from generators, which provide considerably more power than the batteries in the handsets do. Here, the PCS base station has an ERP of 10W at 1 GHz. The antenna height for the base station will be on the order of 10m or so.

12.2 Noise Limits on Detection Range

The amount of noise present at the receiver compared to the amount of signal present determines the detection performance of ES systems. The noise level is situation-dependent. For any kind of reasonable engineering analysis of range performance, the level of noise is either measured or assumed. The SNR, given by

$$\text{SNR} = \frac{P_\text{R}}{N_\text{total}} \tag{12.1}$$

and frequently expressed in decibels, must be above some minimum value in order to detect signals. This level depends on the modulation type and what is trying to be done with the detected signal, but for simplicity here the required SNR level will be 10 dB, which is representative of a broad range of digital modulation types, to include PCS signals. It is, however, pessimistic (too large) for effective collection of analog (AM, FM) modulations [10, 11].

N_total is the total effective noise at the input to the ES system as discussed in Chapter 3. Due to the altitude of the airborne collection systems considered here, it is reasonable to assume that there is little man-made noise to limit their performance. Whereas such noise sources are very important (and limiting) for ground-based military systems embedded with military forces operating equipment that is radiating, such as vehicles, the aircraft are far enough away that these sources are insignificant. Of course, the noise generated by the aircraft itself must be considered but this can usually be filtered out. At least it is known ahead of time. Therefore, the limiting noise source in most airborne cases will be galactic noise. From Figure 3.1 the following noise values are used for this analysis when considering airborne configurations.

$$N_{\text{external}} = \begin{cases} 18 \text{ dB} & f = 30 \text{ MHz} \\ 6 \text{ dB} & f = 100 \text{ MHz} \\ 0 \text{ dB} & f > 200 \text{ MHz} \end{cases} \tag{12.2}$$

When the ES system is ground-based, it is sometimes embedded among many systems that generate man-made noise, such as vehicles and generators. In these circumstances it is more accurate to model the noise environment as suburban man-made limited as in Figure 3.1 rather than galactic noise limited. Thus, N_{external} is given by

$$N_{\text{external}} = \begin{cases} 35 \text{ dB} & 30 \text{ MHz} \\ 24 \text{ dB} & 100 \text{ MHz} \\ 2 \text{ dB} & 1 \text{ GHz} \end{cases} \tag{12.3}$$

12.3 Targets

Three notional target types are considered for comparison purposes here. The first is a 10 W ERP PTT transmitter typical of the low-VHF frequency range common to military ground forces but also typical of PCS base stations, which is the second target type. The third target has typical characteristics of a PCS handset with an ERP of 0.5W at a frequency of 1 GHz. In these cases the transmit antennas are assumed to be omnidirectional.

In Figures 12.2–12.11, the sensor system is placed at $x = 0$ and $y = 0$, and the distances are displayed in kilometers. Since a nonlinear battlespace is the most likely to be encountered by military forces, the detection performance in a 360° sector around the sensor is important.

Military targets in the low-VHF frequency range would typically employ vertical whip antennas, either dipoles or monopoles. The result is predominantly vertical polarization. Ground-based ES system antennas, when stopped and deployed, can readily be built for vertical polarization. Antennas for PCS targets will need to accommodate both polarizations since the handset antennas will be at about any attitude at any time. The base stations will typically be built with vertical polarization.

In addition to these targets, the detection range performance against the sidelobes of two directional antenna types is examined. In particular, two popular directional antennas, the Yagi and the horn-fed parabolic dish, are considered.

12.4 Detection Range with the Reflection Propagation Model

The reflection propagation model described in Section 2.9 given by (2.43) will be used to calculate the signal levels. As previously indicated, at short ranges this model tends to underestimate the amount of power at the receiver. However, at the limits of the intercept range, it is a more accurate model than many others. The limits of the ranges are where this analysis is targeted.

In this chapter, isotropic antennas are assumed for the receiver; thus, $G_R = 1$. As shown previously, actual antennas, optimized for a particular frequency, could have a higher or lower gain than this. Also, reasonable tactical configurations for antennas would not exhibit this gain characteristic. It is assumed here for simplicity and can serve as a basis of departure when actual antenna gains are known.

Since terrain is not considered here, the coverage areas will be circles with the sensor at the center. The radii of the circles will be the calculated detection ranges.

12.4.1 Airborne Configurations of ES Systems

Two airborne configurations are considered. The first is a high-flying system, at an altitude of 10,000m AGL. In the second configuration the aircraft is at 3,000m AGL.

In Figures 12.2–12.11 the SNR required for adequate interception is represented by a horizontal plane in three-dimensional plots. On the other hand, the SNRs will vary depending on where in the target area the target emitter is located. The SNR will also vary with frequency because the noise varies with frequency. Thus, the signal level will be strongest and the SNRs the highest when the targets are near to the sensor system and will be weaker and lower farther away. Anywhere the SNR contours rise above these planes the targets can be detected.

12.4.1.1 High-Flying/Standoff

Results are presented for a single sensor system. Usually two or three such systems are flown at the same time; however, including results in the analysis for more than one platform adds no additional information for detection range determination. Thé coverage area for PF, however, is the conjunction of the coverage regions for two or more sensors. Therefore, in those investigations the overlap of the coverage charts is important.

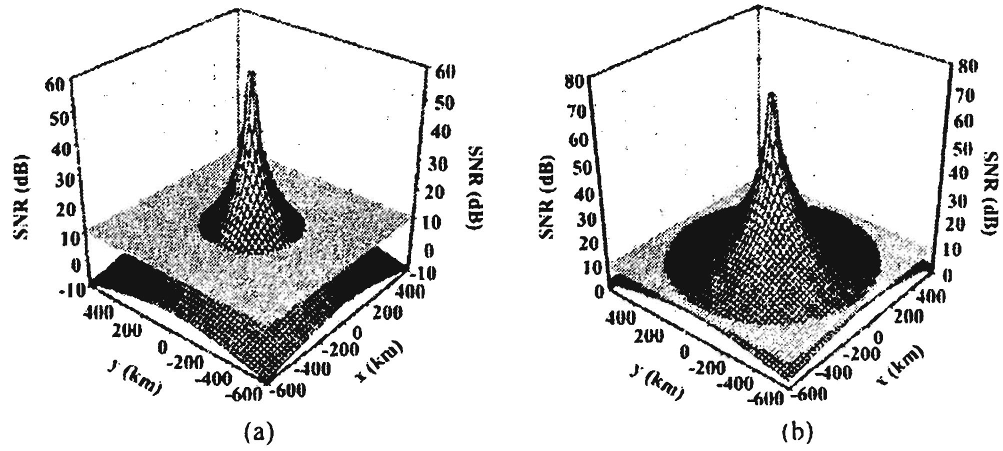


Figure 12.2 Noise-limited range performance for high-flying system: (a) 30 MHz, 10W ERP; and (b) 100 MHz, 10W ERP.

Figure 12.2(a) shows the range performance when the frequency is 30 MHz and the transmitter power is 10W ERP. Note that the geographic scale is ±600 km. The 10 dB SNR signal detection range is approximately 300 km. When the frequency is 100 MHz and the power is 10W ERP, the detection range performance is as shown in Figure 12.2(b). The detection range is larger than when the frequency was 30 MHz. While the signal at a higher frequency does not propagate as well as one at a lower frequency, the noise is less at the higher frequency, yielding greater range performance. Approximately, the RLOS (425 km) is the range over which detection is possible at 100 MHz.

For the PCS frequency range at around 1 GHz, the handset detection performance is shown in Figure 12.3(a) where the ERP is 0.5W. The detection range in this case is about 380 km. Detection of a PCS base station is possible at ranges out to RLOS as shown in Figure 12.3(b).

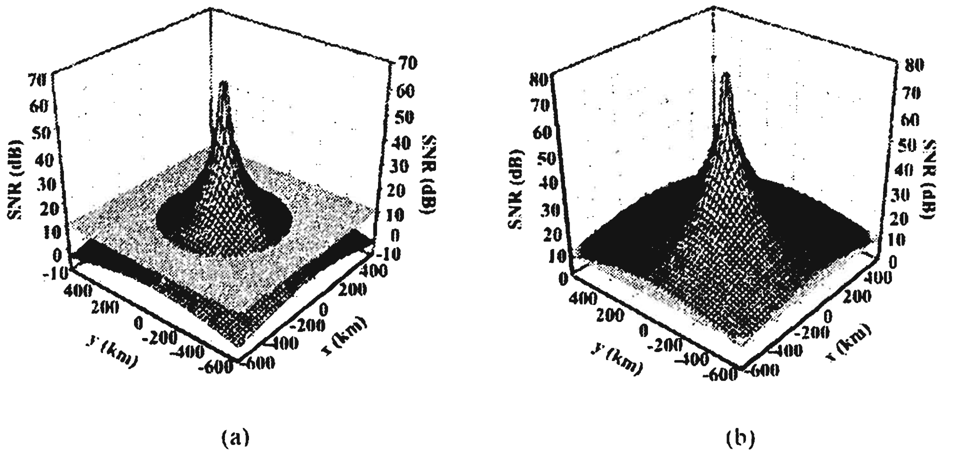


Figure 12.3 Detection range performance of high-flying system: (a) PCS handset at 1 GHz and (b) PCS base station at 1 GHz.

The results presented in Figure 12.3(a) assume that the handset antenna pattern is omnidirectional, thus facilitating equal radiation in all directions. PCS handset antennas have far from omnidirectional antenna patterns, and vary with orientation to the user's body, reflections off local objects such as stop signs, and a myriad of other factors. Unless the target is stationary and in a fairly benign RF environment, the antenna pattern is probably randomly varying. These results, therefore, are optimistic and will not always represent correct detection ranges.

The results in Figure 12.3(b) also assume that the base station antennas are omnidirectional, a poor assumption since PCS base station antennas are typically not. In fact, the expected "smart" base station antennas steer nulls in directions other than toward the particular handset being tracked. Directional antennas at the base station imply that considerably less than 10W ERP will arrive at the second sensor platform even if the first sensor platform is in the correct antenna beam. That would tend to substantially decrease the intercept range of the second platform, degrading the ability to perform real-time PF.

When the high-flying system is attempting to intercept the sidelobes of directional antennas, the range is limited. Using the power levels determined in Section 8.8 for typical Yagi and parabolic dish antennas, and with a 1.5 MHz bandwidth, meaningful at the multichannel frequency ranges considered here, the intercept ranges shown in Figure 12.4 result. In Figure 12.4(a) the target antenna is the Yagi, with an ERP in the direction of the sensor of 0.14W at 1 GHz. Note the scale is ±400 km. The intercept range is approximately 75 km. The target antenna in Figure 12.4(b) is the horn-fed parabolic dish with an ERP of 2 × 10^{-4}W in the direction of the sensor. The scale here is ±60 km. The intercept range is about 24 km. With a standoff high-flying sensor system, neither of these targets would be detected.

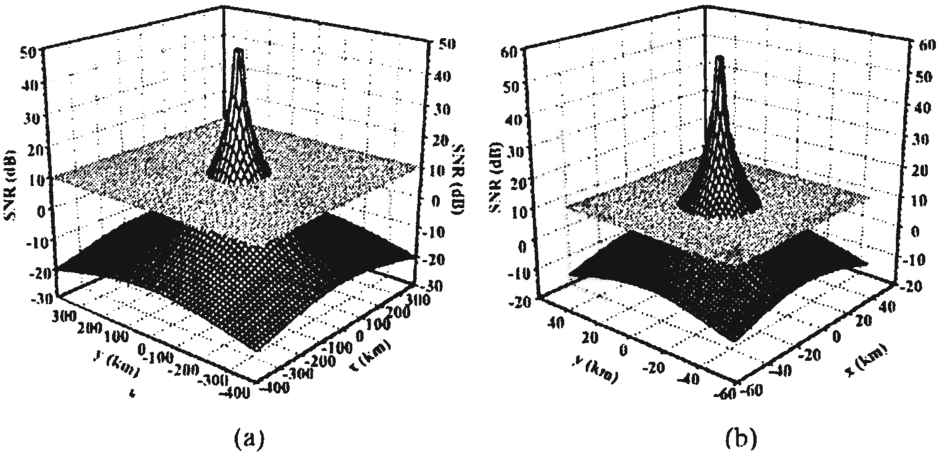
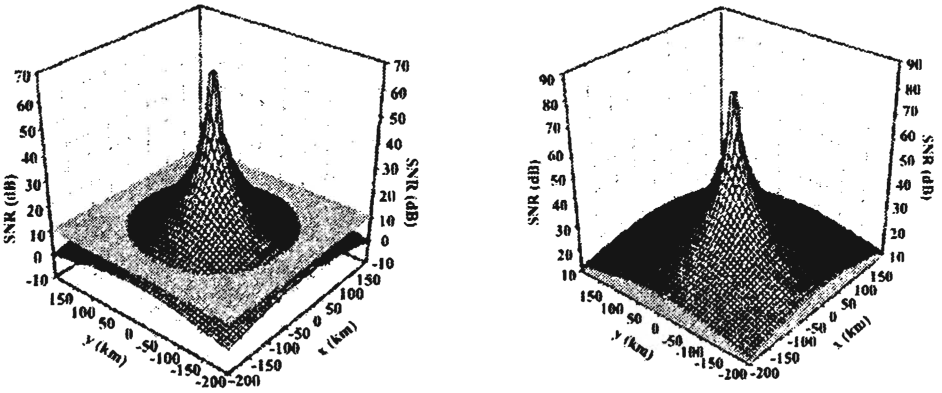


Figure 12.4 Intercept range of a high-flying system when the system is in the sidelobe of a directional antenna: (a) Yagi at 1 GHz and (b) horn-fed parabolic dish at 4 GHz.



(a)

(b)

Figure 12.5 UAS detection range: (a) 30 MHz, 10 W ERP; and (b) 100 MHz, 10 W ERP.

12.4.1.2 UAS

A UAS can be used to augment the collection capability of the high-flying ES system. Such systems have the ability to overfly areas that might be prohibitive for the larger systems for a variety of reasons, some of which were mentioned previously. If the high-flying systems are piloted or otherwise manned, it may be too dangerous in some situations to overfly certain areas. If the system is carrying sensitive equipment for special missions, it may be unwise to risk the loss of such equipment. In mountainous terrain, a UAS can be used to augment the larger system, filling in gaps in the coverage in valleys, for example.

Detection range performance at 30 MHz and 10W ERP is shown in Figure 12.5(a). Note the scale of ±200 km. The SNR is adequate for signal detection to a range of about 160 km or so. The range at 100 MHz and 10W ERP is beyond the RLOS of 200 km as shown in Figure 12.5(b). At 1 GHz, the low-power target at 0.5W ERP can be detected at a range of 200 km or so as shown in Figure 12.6(a). Against the base station, the range increases to beyond RLOS as shown in Figure 12.6(b).

A UAS when overflying a target area has the opportunity to fly into and loiter within the main lobe of a point-to-point directional communication system. This is one of the advantages of a UAS. On the other hand, just as for the higher-flying standoff system, interception of the sidelobe signal is also possible. The intercept range for a UAS flying at 3,000m is shown in Figure 12.7. Figure 12.7(a) is a sidelobe intercept of the Yagi antenna at 1 GHz, with a signal (and noise) bandwidth of 1.5 MHz. The scale in that case is ±60 km and the intercept range is approximately 55 km. Figure 12.7(b) is for sidelobe intercept of the parabolic dish where the scale is ±20 km. The intercept range is about 10 km.

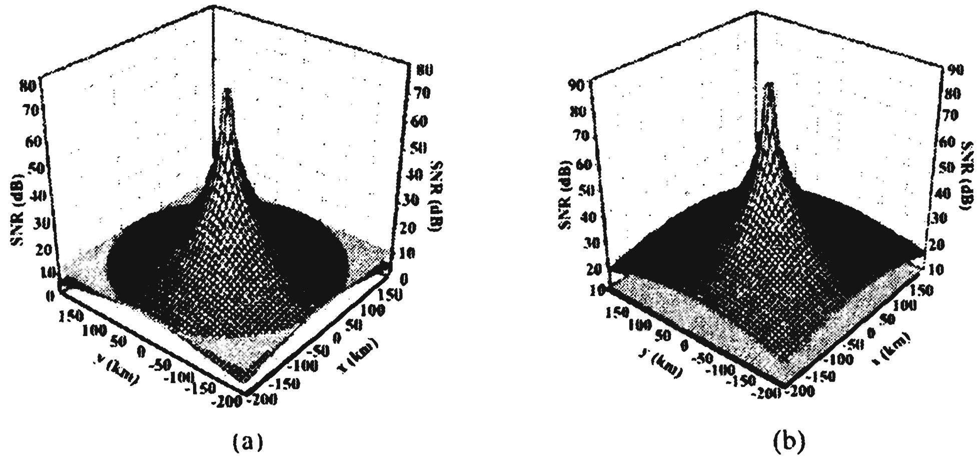


Figure 12.6 Detection range performance from a UAS: (a) PCS handset at 1 GHz and (b) PCS base station target at 1 GHz.

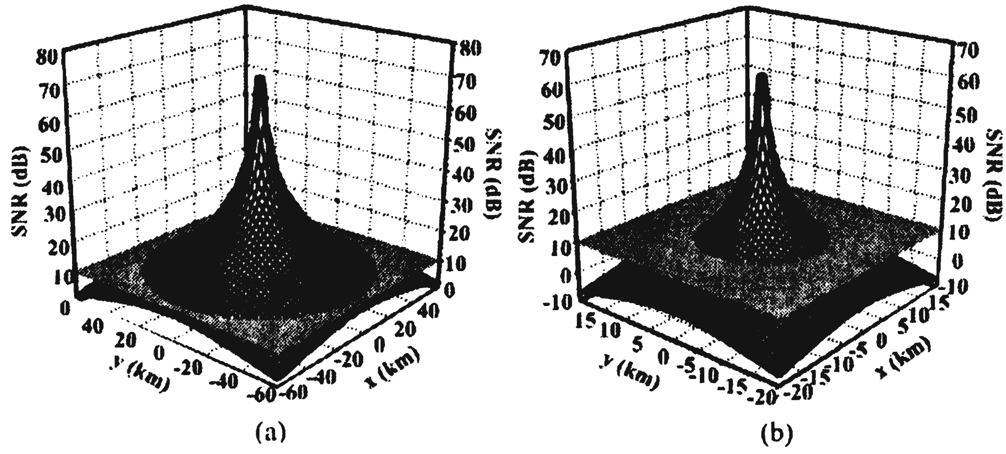


Figure 12.7 Sidelobe intercept range of a UAS: (a) Yagi at 1 GHz and (b) parabolic dish at 4 GHz.

Another implication of these results is that if real-time *position fix* (PF) is desired, two or more aircraft must be flying within the footprints shown. This implies that to cover the gaps in the entire target area, a considerable amount of time may be required, depending on the terrain. This, in turn, implies that a UAS ES system such as this should be used to cover specific areas rather than an entire corps *area of operations* (AO). Of course, real-time PF is not always required, and if delays can be tolerated so that the UAS can be allowed to fly some reasonable distance, then a single aircraft would suffice. GDOP (see Section 11.5) causes this distance to be on the order of the same as the range to the target, 5–10 km in this case, in order to produce reasonably accurate locations. At 100 km per hour aircraft speed this may create collection problems, as the target may go away in the 3 minutes required to fly 10 km.

Extensive range against PCS cell base stations and handsets implies that the sensitivity of the ES system must be decreased in order to minimize cochannel interference. The sensitivity should be set so that no more than one set of the seven cells making up a group in Figure 12.1 are covered at a time. A typical cell radius is 10 km.

Likewise, the extensive intercept range at low-VHF could require decreasing the sensitivity there as well. Since low-VHF frequency-hopping targets utilize much of the spectrum, these intercept ranges could cause cochannel interference, especially since the spectrum is shared by red, blue, and gray forces/populations.

12.4.2 Ground-Based Detection Ranges

Two ground ES system configurations are considered. The first has an elevated antenna on a short mast at h_R = 10m reflecting the performance of a stationary deployment. The second antenna height is h_R = 3m, reflecting expected performance while OTM.

As shown in Figure 12.8, at 30 MHz and 100 MHz, the detection range is about 3 km and 6 km, respectively, from the sensor for a target ERP of 10W and h_R = 10m. The scale in these figures is ±20 km. The limited range is because of the excessively high noise levels at these low frequencies in the suburban man-made noise environment.

This noise environment would not be encountered in all cases. It is probably appropriate for road marches or when there is other significant battlespace activity. When preparing for hostile action, but prior to it, a circumstance where ES systems are normally employed is that the noise would be substantially less, especially at night. Nevertheless, this analysis points out that the range of ground-based communication ES systems is very limited.

In the higher-frequency range around 1 GHz, there is insignificant external noise impinging on the receive antenna (about 2 dB_{kTB}). Thus the noise effects are

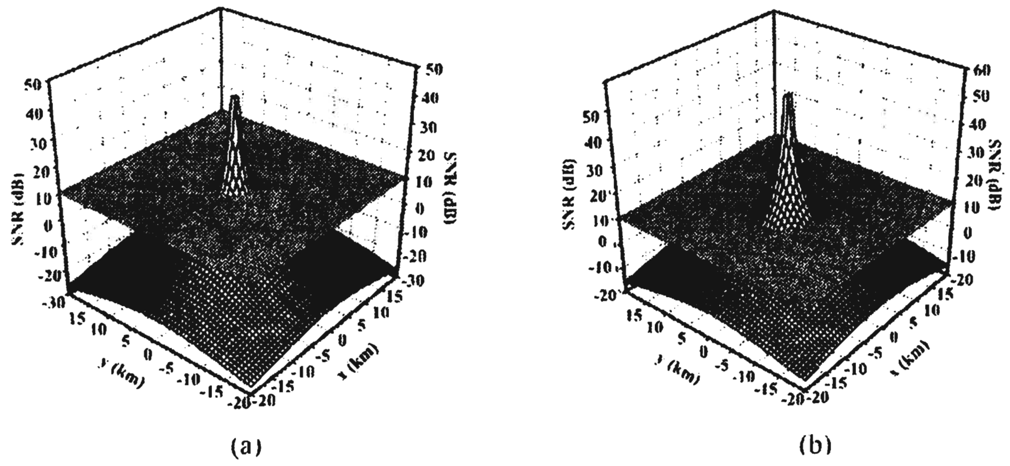


Figure 12.8 Ground-based detection range for h_R = 10m: (a) 30 MHz, 10W ERP; and (b) 100 MHz, 10W ERP.

small compared to the low VHF. These effects are shown in Figure 12.9. The handset range is about 10 Km and the base station is out to RLOS (16 km). One of the principal functions of communication ES is to support operations OTM. When moving, the antenna height cannot be too high. These results assume that the receive antenna is erected just above the vehicle, at a height from the ground of 3m. The results are shown in Figure 12.10(a) for a target frequency of 30 MHz and a target ERP of 10W. The scale here is ±10 km. The detection range is about 2 km. At 100 MHz, where the man-made noise is less, the detection range, from Figure 12.10(b), is about 3 km.

OTM performance against the PCS system is shown in Figure 12.11. Against the 1 GHz, 0.5W ERP handset the detection range is about 6 km, reflecting the relatively low external noise at this frequency (2 dB_{kTB}). Against the 10W ERP base station the detection range is beyond RLOS.

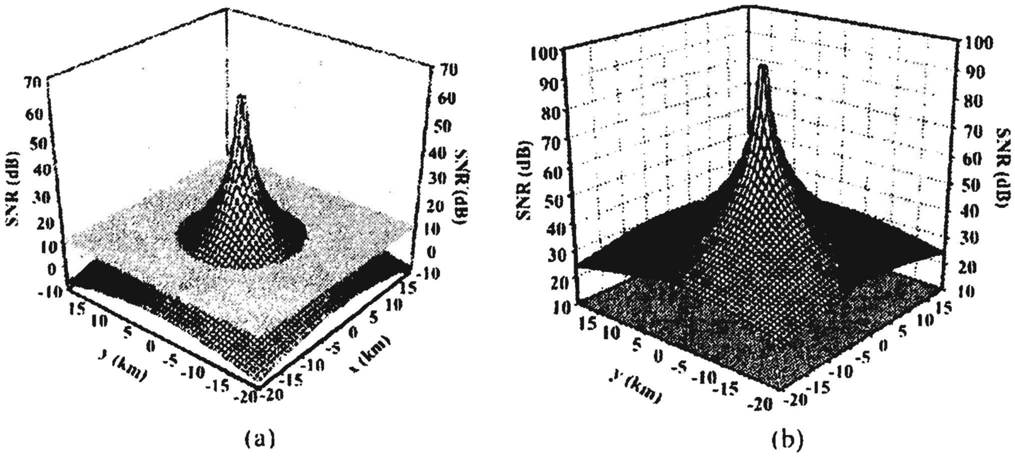
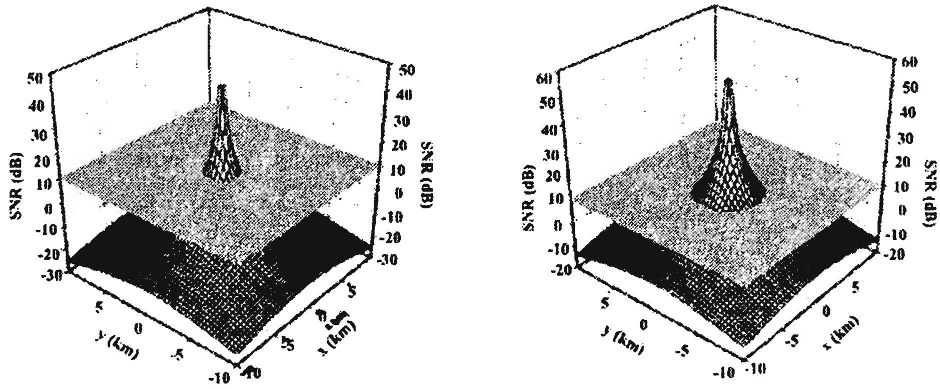
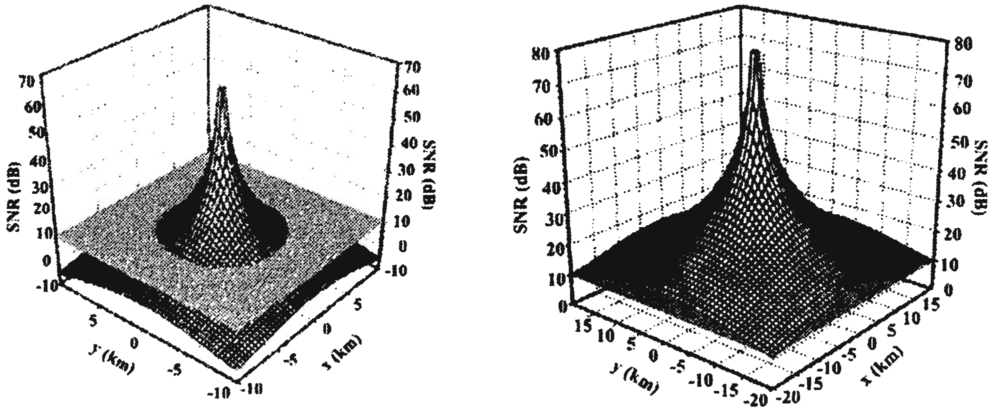


Figure 12.9 Ground detection range for h_R = 10m: (a) PCS handset at 1 GHz with a 0.5W ERP; and (b) PCS base station at 1 GHz with an ERP of 10W.



(a) (b)

Figure 12.10 Ground based detection range while OTM when h_R = 3m: (a) 30 MHz, 10W ERP; and (b) 100 MHz, 10W ERP.



(a) (b)

Figure 12.11 Ground detection range while OTM when h_R = 3m: (a) PCS handset at 1 GHz, 0.5W ERP; and (b) PCS base station at 1 GHz, 10W ERP.

Table 12.1 Detection Ranges for the System Configurations Considered (km)

	Target			
	0.5W, 1 GHz	**10W, 30 MHz**	**10W, 100 MHz**	**10W, 1 GHz**
High-flying	380	300	RLOS (425)	RLOS (425)
Low-flying	200	160	RLOS (200)	RLOS (200)
Ground	10	3	6	RLOS (18)

12.4.3 Discussion

Generally speaking, the condition that must be met for signal detection is that the signal must have a certain SNR or higher at the receiver. When external noise is not considered, an E-field level can equivalently specify that SNR. The detection ranges concluded from this analysis are summarized in Table 12.1.

Because in the overflight configuration both the high-flying and low-flying configurations cover a significant portion of the low-VHF band, near-optimum geometries are possible for PF. The best geometry for CAF TDOA and DD PF processing (see Section 12.4) has the targets more or less between the sensor systems. This can be accomplished by flying a UAS on the other side of the primary target area, placing the targets for which precision locations are required between the sensors. Not only is the geometry near optimum this way, but also the SNR is higher at the UAS, which can substantially increase the system throughput by decreasing the processing time necessary to compute the locations.

If CAF processing is too slow or not available, a PF geometry consisting of an L-shaped configuration is much better than PF with a linear baseline. The optimal geometry for DF processing is to have the LOPs cross at 90° angles. This occurs more frequently for L-shaped baselines than for linear baselines. An L-shaped baseline is available with two high-flying sensors on a linear baseline with a UAS on the other side or to the rear of the primary target area.

Intercept range performance by detection in directional link sidelobes is sometimes possible with either airborne system. In Table 12.2 the range performance is summarized. With these ranges, overflight of the target area is required.

Ground systems have limited range, limited by noise as well as RLOS. The latter of these should be applied carefully, however. On the ground, as opposed to the air, detection range is sensitive to system elevation relative to surrounding terrain. Both air and ground configurations, however, must be concerned with

Table 12.2 Sidelobe Detection Range (km)

	Yagi	**Parabolic Dish**
High-flying standoff	75	24
UAS	55	10

P_T = 10W, Yagi at 1 GHz and parabolic dish at 4 GHz

Table 12.3 Ground-Only Detection Ranges (km)

	0.5W, 1 GHz	10W, 30 MHz	10W, 100 MHz	10W, 1 GHz
Stationary	10	3	6	RLOS (18)
OTM	6	2	3	RLOS (12)

signal blockage. In OTM operations the RLOS is about 12 km over smooth Earth. Over hilly terrain or roads, the detection range would be limited more by blockage and temporary elevation advantages than detection range limits imposed by reflections and RLOS. In general, this is true for any ground deployment. These results are summarized in Table 12.3, which compares the stationary versus OTM detection ranges.

12.5 Concluding Remarks

This chapter presented some expected performance information on airborne and ground-based ES systems. Two airborne configurations are considered—one at a high altitude and one at a lower altitude. The flight characteristics of the latter are compatible with those expected with a UAS. Two ground system configurations were also examined—one stationary and one OTM.

The analysis presented herein indicates that augmenting the high-flying ES collection system with a low-flying UAS can in many cases improve the collection performance. Indeed, in some scenarios the UAS could be the only alternative—in circumstances where it is desired to not use the high-flying system for safety or other reasons. Either system can cover the corps AO if there are no terrain blockages as shown in Figure 12.12. With terrain blockage, the UAS can overfly such terrain to obtain the targets. The UAS has enough intercept range to detect signals in these coverage zones. The UAS also has sufficient range to augment the DF/CAF processing of the high-flying system. When the targets are PCS/cell communication systems, the coverage is extensive enough from either airborne system so that the sensitivity must be decreased to avoid hearing multiple cells at the same time.

There are clear scenarios where not only is a UAS augmentation to the high-flying ES system desirable, but it is also required. In low-power and directional target situations the high-flying system simply cannot detect the targets without flying very close to them. For these targets, a standoff configuration cannot detect them.

The high-flying systems, being manned platforms, are vulnerable to enemy countermeasures. Since it is possible to perform ES collection with unmanned platforms, that should be the method of choice.

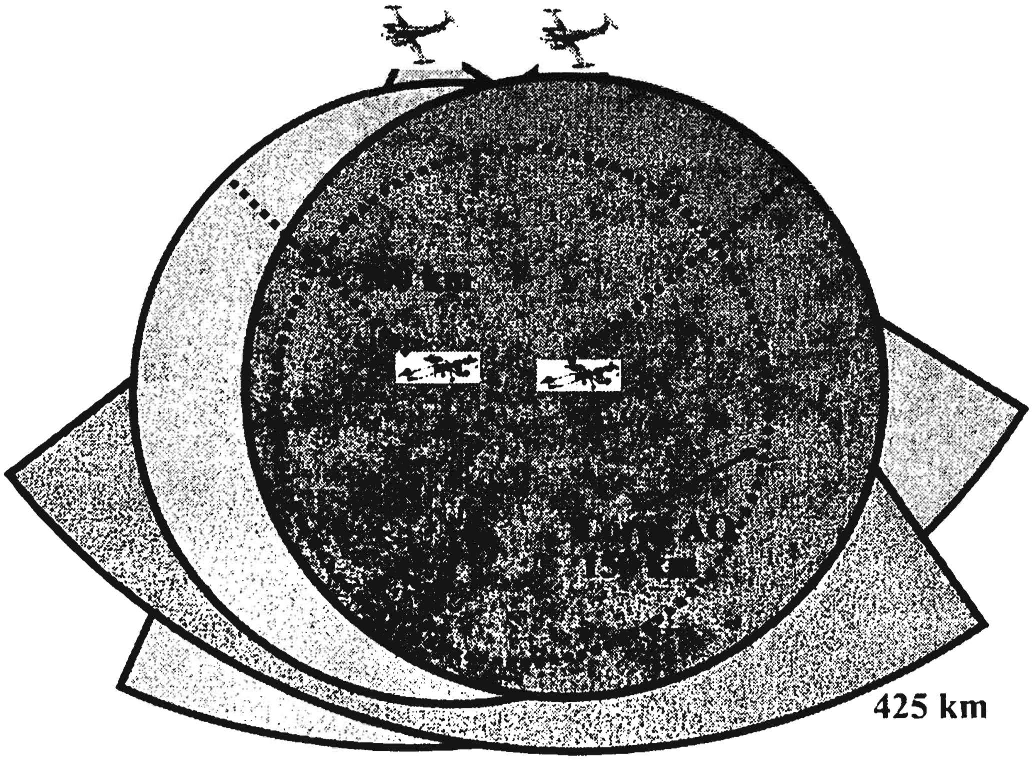


Figure 12.12 Coverage zones of a corps AO with both high-flying standoff systems augmented with overflying UASs.

There are additional situations where the geometry provided by a UAS augmentation facilitates significantly enhanced PF of targets. Keeping a target between two airborne platforms is the optimum geometry for CAF computations. Furthermore, the SNRs available in that situation improve system throughput significantly.

A scenario where a UAS may be the only choice is when ground access is not feasible. A ground-based PCS ES system is a reasonable alternative to airborne solutions if access is available. The ground equipment must be placed somehow and may have to be guarded. If security is not an issue, then the ground assets can be air-dropped close to base stations and remotely controlled and monitored. If there are security concerns, then manning the ground assets is probably required, even if collection is remote. A much simpler solution to this problem is the UAS ES system.

The UAS should be employed as a precision sensor system as opposed to a general search system. Their largest advantage is to overfly the target area picking up low-power and masked targets. In order to do PF, this will normally require at least two platforms. They must fly fairly close together to simultaneously detect low-power and masked targets.

One of the advantages of the UAS augmentation is that in some scenarios, in particular those where the high-flying system is in a standoff, rear deployment, the UAS sensor need not look through and sort through many friendly emitters to get to the targets. An emission in this case must first be considered friendly especially if it is stronger. The UAS can fly among the targets in this case, where every emission is likely to be a target as opposed to friendly. The coverage of high-flying standoff systems augmented with UASs is shown in Figure 12.12. Both the standoff systems and the UASs have adequate range to cover the entire AO.

When the targets are directional point-to-point communication links, the high-flying system in a standoff posture has difficulties with intercept range. If the communication system uses Yagi antennas, there is insufficient leakage into the sidelobes to intercept these systems at an adequate distance into the corps AO. Against such links that are using parabolic dishes, there is also inadequate leakage into the sidelobes to intercept them from a standoff posture. A UAS, on the other hand, can either fly into the main beam and loiter there if required, or intercept either of these links in a sidelobe by getting close enough.

This analysis considered whether targets could be detected versus range from the sensor system. It did not consider timing concerns, such as system throughput, or timing considerations associated with geolocating or collecting LPI targets.

References

[1] Shigekazu, S., *A Basic Atlas of Radio-Wave Propagation*, New York: Wiley, 1987.

[2] Hall, M. P. M., *Effects of the Troposphere on Radio Communications*, London: Peter Peregrinus, Ltd., 1979.

[3] Braun, G., *Planning and Engineering of Shortwave Links*, 2nd ed., New York: Wiley, 1986.

[4] Parsons, J. D., *The Mobile Radio Propagation Channel*, New York: Wiley, 1992.

[5] Volland, H. (ed.), *Handbook of Atmospherics*, Vols. I and II, Boca Raton, FL: CRC Press, 1982.

[6] Davies, K., *Ionospheric Radio*, London: Peter Peregrinus, Ltd., 1989.

[7] Schnker, J. Z., *Meteor Burst Communications*, Norwood, MA: Artech House, 1990.

[8] *Reference Data for Radio Engineers*, 6th ed., Indianapolis, IN: Howard W. Sams & Co., 1975.

[9] Calhoun, G., *Digital Cellular Radio*, Norwood, MA: Artech House, 1988.

[10] Gagliardi, R. M., *Introduction to Communications Engineering*, New York: Wiley, 1988.

[11] *Reference Data for Radio Engineers*, 6th ed., Indianapolis, IN: Howard W. Sams & Co., 1975, Ch. 29.

Chapter 13

Jamming Performance in Fading Channels

13.1 Introduction

Up until now, it has been assumed that the signaling between the target network transmitter and receiver as well as the jammer signal at the receiver do not fade. As discussed in Chapter 2, however, almost all realistic communication links do indeed fade, primarily caused by the presence of several multipath channels. This fading has significant affects on the communication link performance as well as the jammer–link performance as will be discussed in this chapter. This chapter presents the effects of signal fading on jammer performance. These signals can be target signals or jamming signals—the fading affects both (although that's not to say that the requirement to counter such fading is the same).

The effects on jamming performance when only the target signal fades, as well as when both the target signal and the jamming signal fades, are included. The type of fading on the target signal and jamming signal need not be the same since frequently different propagation channels are involved.

This chapter is structured as follows. We begin in the next section with analysis of signal detection probability when the channel exhibits fading. That is followed by a section where jamming performance against narrowband targets is examined, also when the channels fade. Next, jamming performance is evaluated when the targets use DSSS signaling and the channels fade. After that, the targets are assumed to be FHSS with fading channels. The last section considers the case when the targets are hybrids of DSSS and FHSS.

13.2 Probability of Detection in Fading Channels

The EW problem of signal detection differs from that of the target networks. The target transmitters can cooperate with the receivers in a variety of ways in order to

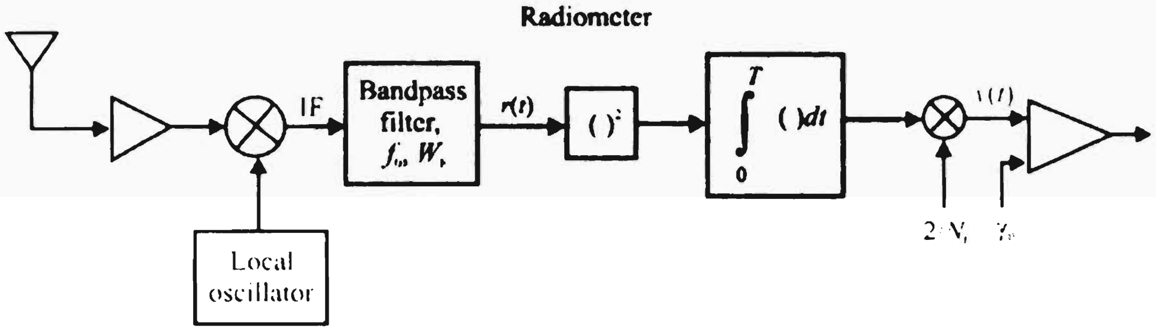


Figure 13.1 Radiometer.

improve communication performance. The EW system in general cannot do that. One means of cooperation is to implement diversity reception where the multipath channels can be estimated and compensated for. Much of the literature on detection in fading channels discusses that topic. We will not discuss diversity reception further here because it does not apply to our problem.

Urkowitz was the first to examine the detection performance of unknown deterministic signals over a flat band-limited Gaussian noise channel [1]. The receiver is depicted in Figure 13.1, which is a radiometer. The energy in the received signal, shown as an IF signal in Figure 13.1, is measured over an observation time window T. After T seconds the integrated signal is compared to a threshold, γ_{th}, and if the result is larger than the threshold, then detection is declared. Kostylev [2] examined this problem including Rayleigh, Rice, and Nakagami fading. Fading is caused by receiving several versions of the transmitted signal over separate paths. Each path imparts its own attenuation and phase shift and when the signals are linearly combined at the receiver, degradation of the total signal occurs. Furthermore, if any portion of any of the channels is in motion, the fading will be nonstationary. Such movement need not be the transmitter or receiver, but could be a reflecting surface in the path, for example, or a large truck moving.

13.2.1 AWGN Channels

Denote the two signal detection hypotheses as

- H_0: signal absent: $y(t) = n(t)$
- H_1: signal present: $y(t) = \xi s(t) + n(t)$

With the radiometer, the probabilities of detection, defined as

$$P_d = \Pr\{y > \gamma_{th} | H_1\} \quad (13.1)$$

and false alarm, defined as

$$P_{\text{fa}} = \Pr\{y > \gamma_{\text{th}} | H_0\} \tag{13.2}$$

over flat (nonfading) AWGN channels are given by [3]

$$P_{\text{d}} = Q_{N/2}\left(\sqrt{\frac{\mu\gamma}{\sigma^2}}, \sqrt{\frac{\gamma_{\text{th}}}{\sigma^2}}\right) \tag{13.3}$$

and

$$P_{\text{fa}} = \frac{\Gamma\left(\frac{N}{2}, \frac{\gamma_{\text{th}}}{2\sigma^2}\right)}{\Gamma\left(\frac{N}{2}\right)} \tag{13.4}$$

respectively, where

- $Q_K(\cdot,\cdot)$ is the generalized Marcum Q-function[1]
- $\Gamma(\cdot\,,\cdot)$ is the incomplete gamma function
- $\gamma = \alpha^2 \frac{E_\text{b}}{N_0}$ is the SNR
- $\xi = \alpha e^{j\theta}$ is the channel slow fading characteristic (without fading $\alpha = 1$ and $\theta = 0$)
- $\mu = a\gamma$ for some positive constant a
- $N \approx TW$ = number of samples = number of degrees of freedom
- W = bandwidth

As pointed out in Chapter 7, when P_d (or equivalently the probability of miss, $P_\text{m} = 1 - P_\text{d}$) and P_fa are plotted against each other, the receiver function that is displayed is referred to as a receiver operating characteristic, or ROC, an example of which is illustrated in Figure 13.2.

[1] The generalized mth order Marcum Q-function is given by

$$Q_m(a,b) = \frac{1}{a^{m-1}} \int_b^\infty x^m e^{-\frac{x^2+a^2}{2}} I_{m-1}(ax)dx$$

where $I_n(.)$ is the nth order Bessel function of the first kind.

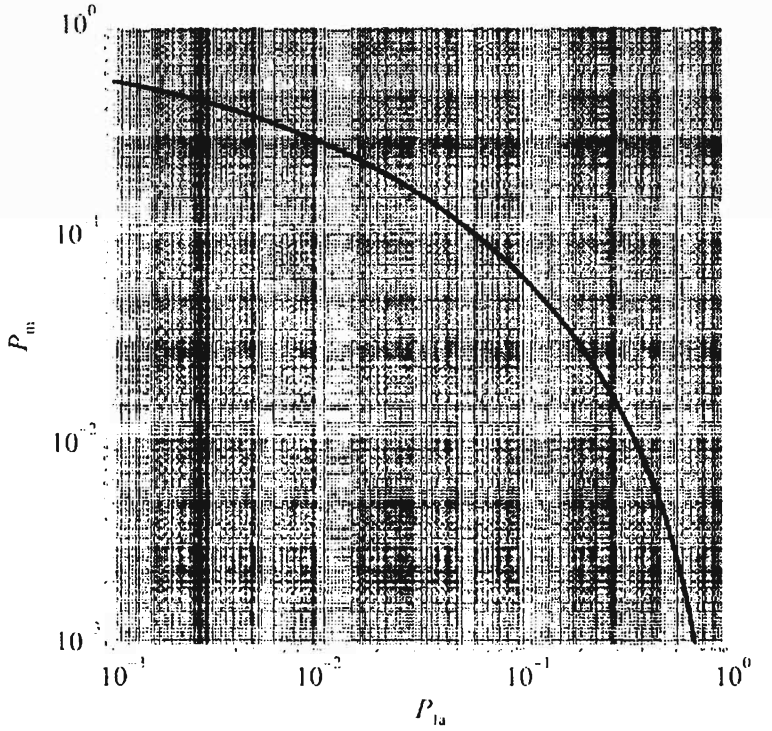


Figure 13.2 ROC without fading. $N = 10$ ($\gamma = 10$ dB, $\mu = 2$, $\sigma = 1$).

13.2.2 Probability of Detection over Nakagami and Rayleigh Fading Channels

The average detection probability in Nakagami and Rayleigh fading channels is discussed in this section. The probability of false alarm, (13.4), will not change due to fading because the false alarm rate is independent of γ. The detection probability for a Nakagami channel is given by [3]

$$P_{\mathrm{d,Nak}} = A_1 + \beta^m e^{-\gamma_{\mathrm{th}}/2\sigma^2} \sum_{i=1}^{N/2-1} \frac{\left(\frac{\gamma_{th}}{2\sigma^2}\right)^i}{i!} {}_1F_1\left[m; i+1; \frac{\gamma_{\mathrm{th}}(1-\beta)}{2\sigma^2}\right] \tag{13.5}$$

where ${}_1F_1(\cdot;\cdot;\cdot)$ is the *confluent hypergeometric function*,[2] and m is *the Nakagami parameter*. For integer m

[2] The confluent hypergeometric function is given by

$${}_1F_1(a;b;z) = \sum_{k=0}^{\infty} \frac{(a)_k}{(b)_k} \frac{z^k}{k!}$$

where $(a)_k = a(a+1)(a+2)\cdots(a+k-1)$, $(a)_0 = 1$.

$$A_1 = e^{-\gamma_{\text{th}}\beta/2m\sigma^2}\left\{\beta^{m-1}L_{m-1}\left[-\frac{\gamma_{\text{th}}(1-\beta)}{2\sigma^2}\right]+(1-\beta)\sum_{i=0}^{m-2}\beta^i L_i\left[-\frac{\gamma_{\text{th}}(1-\beta)}{2\sigma^2}\right]\right\} \tag{13.6}$$

where $L_i(\cdot)$ is the *Laguerre polynomial*[3] [3] of degree i. Parameter β is given by

$$\beta = \frac{2m\sigma^2}{2m\sigma^2 + a\bar{\gamma}} \tag{13.7}$$

13.2.2.1 Rayleigh Fading Channel

$P_{\text{d,Ray}}$ can be obtained from (13.5) by setting $m = 1$ yielding

$$\begin{aligned} P_{\text{d,Ray}} = e^{-\frac{\gamma_{\text{th}}}{2\sigma^2}}\sum_{i=0}^{N/2-2}\frac{\left(\frac{\gamma_{\text{th}}}{2\sigma^2}\right)^i}{i!} &+ \left(\frac{2\sigma^2+\mu\bar{\gamma}}{\mu\gamma}\right)^{N/2-1} \\ &\times\left[e^{-\frac{\gamma_{\text{th}}}{2\sigma^2+\mu\gamma}} - e^{-\frac{\gamma_{\text{th}}}{2\sigma^2}}\sum_{i=0}^{N/2-2}\frac{\left[\frac{\gamma_{\text{th}}\mu\bar{\gamma}}{2\sigma^2(2\sigma^2+\mu\bar{\gamma})}\right]^i}{i!}\right] \end{aligned} \tag{13.8}$$

A representative ROC corresponding to (13.8) is illustrated in Figure 13.3 where $\bar{\gamma}$ is varied from 10 dB to 20 dB, typical values.

13.2.2.2 Ricean Fading Channel

The average probability of detection without diversity when the channel fading is characterized as Ricean is given by [4]

[3] The Laguerre polynomial $L_i(z)$ can be defined by the contour integral

$$L_i(z) = \frac{1}{2\pi j}\oint\frac{e^{zt/1-t}}{(1-t)t^{i+1}}dt$$

where the contour encloses the origin and is traversed in a counterclockwise direction.

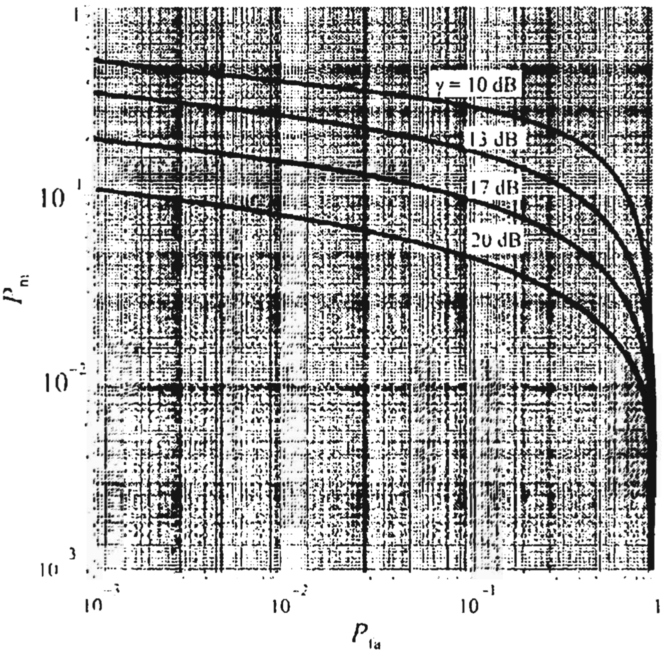


Figure 13.3 ROC curves for Rayleigh fading with various SNRs (N = 10, a = 2, σ = 1).

$$P_{\text{d,Ric}|u=1} = Q\left(\sqrt{\frac{2K\bar{\gamma}}{K+1+\bar{\gamma}}}, \sqrt{\frac{\gamma_{\text{th}}(K+1)}{K+1+\bar{\gamma}}}\right) \tag{13.9}$$

where K is the Ricean K-factor and $u = TW_F = 1$. As K is varied, the curves illustrated in Figure 13.4 ensue. (Note that $N = 10$ in Figure 13.3 while $u = TW_F = N/2 = 1$ in Figure 13.4.) $K = 0$ corresponds to Rayleigh fading since K is defined as the ratio of the direct wave power to the sum of the power in all reflected components, and $K = 0$ implies that there is no direct wave. As the power in the direct wave is increased, the detection performance improves as shown in Figure 13.4.

13.3 Jamming Fading Narrowband Targets

The SNR present at the input to the target receiver is an important system parameter. As previously indicated, in an analog communication system, the SNR is directly related to the subjective quantities such as the intelligibility of the voice or video. In digital systems, as discussed in Chapter 4, the BER is determined by this SNR once the modulation is specified.

Fading implies time variability of the fading signals, whether the signal is from the transmitter or the jammer. Such fading causes γ to vary with time. The

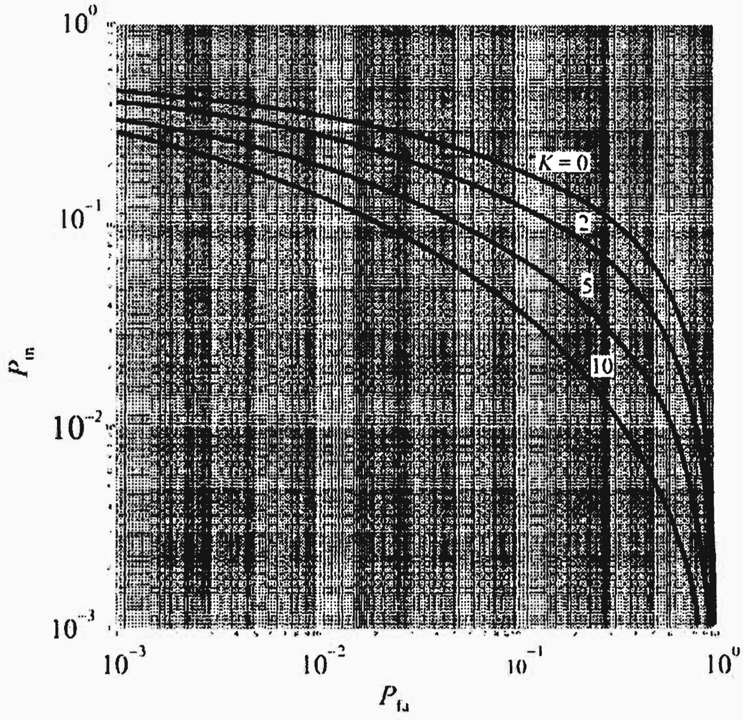


Figure 13.4 Ricean ROC K factor (γ = 10 dB, u = 1).

fading may be caused by movement of the transmitter or receiver in and out of shadowing from large structures such as mountains or buildings. It may also be caused by varying multipath conditions also caused by movement of the transmitter or receiver, or other objects on one of the paths from the transmitter to the receiver. Essentially all ground-to-ground RF communication links experience fading as do all skywave HF links, troposcatter links, ducting links, underwater acoustic links, and satellite links.

Noise, on the other hand, can frequently be modeled as time-stationary due to its origins. Internal thermal noise is predominant at the higher frequencies while atmospheric, man-made, and cosmic noise dominate at the lower frequencies. Therefore, variation in γ is caused almost exclusively by fading. As indicated in Chapter 4, if the signal fading is caused by a large number of diffuse components with no significant main signal path or component, then the envelope of the received signal is approximated by a Rayleigh distribution. On the other hand, if there is a significant direct component to the signal, and there is a nonfading component, then it is modeled by a Ricean distribution.

If the noise is dominated by a jammer, then both the signal as well as the jamming signal may be subject to fading. Furthermore, this fading need not be of the same type, depending on whether there is a strong nonspecular component to the signal, the jammer signal, or both. If they are both Ricean, then ξ = JSR = J / S

is the ratio of two Ricean random variables. Likewise, if there is no nonspecular component to either of the signals then ξ is a ratio of two Rayleigh random variables. Finally, if one is Ricean and the other Rayleigh, then ξ is mixed. The Rayleigh distribution, however, is a special case of the Ricean distribution, obtained by setting the amplitude of the specular component to zero. The Ricean/Rayleigh, Rayleigh/Ricean, and Rayleigh/Rayleigh cases can be obtained from the more general Ricean/Ricean result. This distribution is presented in this section [5].

Most *antijam* (AJ) communication systems exhibit a threshold effect; that is, there is a threshold SNR below which the system does not perform well and the BER rises significantly. The same is true when the noise is intentionally injected as from a jammer. Therefore, the degree by which an AJ system is effective against jamming can be indicated by the probability that this threshold is exceeded. Likewise, the effectiveness of EW measures against such AJ systems can be indicated by the probability that the threshold is not met. The measure given by the probability that the threshold is exceeded is called the *availability*.

13.3.1 Jamming Narrowband Targets

Jamming of narrowband systems in fading channels is considered first. Narrowband systems in this case refers to the classical definition of the term. The effects on wideband (FHSS and DSSS) systems are considered later. The multipath effects on narrowband systems can be different from that for wideband systems because of the effects imposed by the coherence time of the channel. By definition, the wideband systems incorporate considerably more frequencies and the fading characteristics are typically frequency dependent. Denoting the SJR = γ_J and the JSR = ξ, then $\gamma_J = 1/\xi$.

The pdf of the ratio of two signals $z = x / y$ with Ricean fading amplitudes is[4]

$$f_Z(z) = \frac{2c^2 z}{(z^2+c^2)^2} \exp\left(-\frac{c^2 r_S^2 + z^2 r_J^2}{z^2+c^2}\right) \times \left[\left(1+\frac{z^2 r_S^2 + c^2 r_J^2}{z^2+c^2}\right) I_0(\alpha) + \alpha I_1(\alpha)\right], \qquad z \geq 0 \tag{13.10}$$

where $I_0(\cdot)$ is the modified Bessel function of the first kind of order zero. r_s^2 is the ratio of the power of the specular component to the diffuse component for the target signal while r_J^2 is the same for the jammer signal. Thus

[4] This description is a summary of the derivation in [5].

$$r_\mathrm{S}^2 = \frac{a_\mathrm{S}^2}{2\sigma_\mathrm{S}^2} \tag{13.11}$$

$$r_\mathrm{J}^2 = \frac{a_\mathrm{J}^2}{2\sigma_\mathrm{J}^2} \tag{13.12}$$

The target signal is Ricean with parameters σ_S^2 and a_S while the jammer signal is Ricean with parameters σ_J^2 and a_J; c^2 is the ratio of the variance of the target signal's amplitude to the variance of the jammer's signal amplitude at the target receiver

$$c^2 = \frac{\sigma_\mathrm{S}^2}{\sigma_\mathrm{J}^2} \tag{13.13}$$

and

$$\alpha(z) = \frac{2cr_\mathrm{S}r_\mathrm{J}z}{z^2+c^2} \tag{13.14}$$

If the fading is Ricean/Ricean then (13.10) applies unmodified. If there is no significant direct component from the jammer but there is from the target transmitter then the fading is Ricean/Rayleigh and the pdf of γ_J is determined by setting $r_\mathrm{J} = 0$ in (13.14) and (13.10), since $\alpha = 0$, and $\mathrm{I}_0(0) = 1$ to obtain

$$f_Z(z) = \frac{2c^2z}{(z^2+c^2)^2}\exp\left(-\frac{c^2r_\mathrm{S}^2}{z^2+c^2}\right)\left[\left(1+\frac{z^2r_\mathrm{S}^2}{z^2+c^2}\right)\right], \qquad z \geq 0 \tag{13.15}$$

On the other hand, if there is no significant direct component from the target transmitter, but there is from the jammer, then the pdf is Rayleigh/Ricean, and is determined by setting $r_\mathrm{S} = 0$ in (13.14) and (13.10) to obtain

$$f_Z(z) = \frac{2c^2z}{(z^2+c^2)^2}\exp\left(-\frac{z^2r_\mathrm{J}^2}{z^2+c^2}\right)\left[\left(1+\frac{c^2r_\mathrm{J}^2}{z^2+c^2}\right)\right], \tag{13.16}$$

Finally, if there is no significant direct component at the receiver from either the target transmitter or the jammer then the fading is Rayleigh/Rayleigh and the pdf of γ_J is given by

$$f_Z(z) = \frac{2c^2 z}{(z^2 + c^2)^2}, \qquad z \geq 0 \tag{13.17}$$

13.3.1.1 Slow Fading

Consider the case of slow fading, where γ_J has no time variability within the same message but can vary from message to message. The probability of jamming is used as the jammer MOE in the presence of fading. If γ_{th} represents the threshold above which acceptable communication ensues, then the average probability of jamming can be determined by integrating (13.10) or one of the simplifications depending on the type of fading present, and subtracting that value from one:

$$\bar{P}_J = 1 - \int_{\sqrt{\gamma_{th}}}^{\infty} f_Z(z)dz \tag{13.18}$$

To achieve a 10^{-1} BER for noncoherent BFSK, Figure 13.5 [6] indicates that an unfading SNR of 5 dB is adequate. Therefore it is a reasonable goal for a jammer to try to achieve an SJR of 5 dB or worse in the channel to provide the same BER. The probability of jamming versus ξ at this threshold value is illustrated in Figure 13.6 for the fading characterization being considered here.

13.3.1.2 Faster Fading

Next we consider moderately slow fading where the ξ does vary within the same message. When the fading is slow compared to the data rate but fast compared to the message duration, the average probability of error can be used as the measure of jamming performance. This probability of error can be found by using the expression for the probability of error for the nonfading case and averaging over all possible values of γ. The same procedure can be applied to the situation in which the probability of error depends on the random variable γ_J. For example, the probability of error.for nonfading, noncoherent detection of BFSK is given by

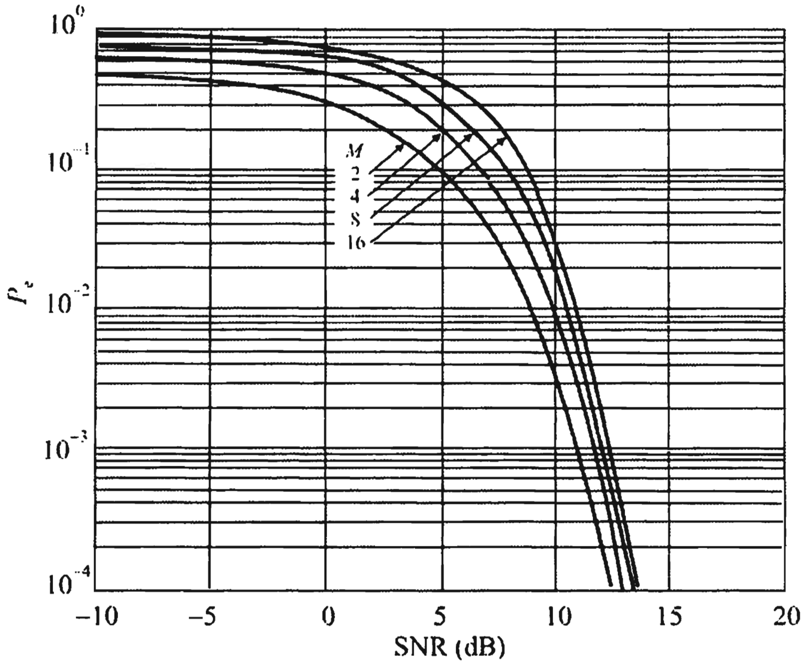


Figure 13.5 MFSK performance without fading. (*Source:* [6]. © Artech House 2004. Reprinted with permission.)

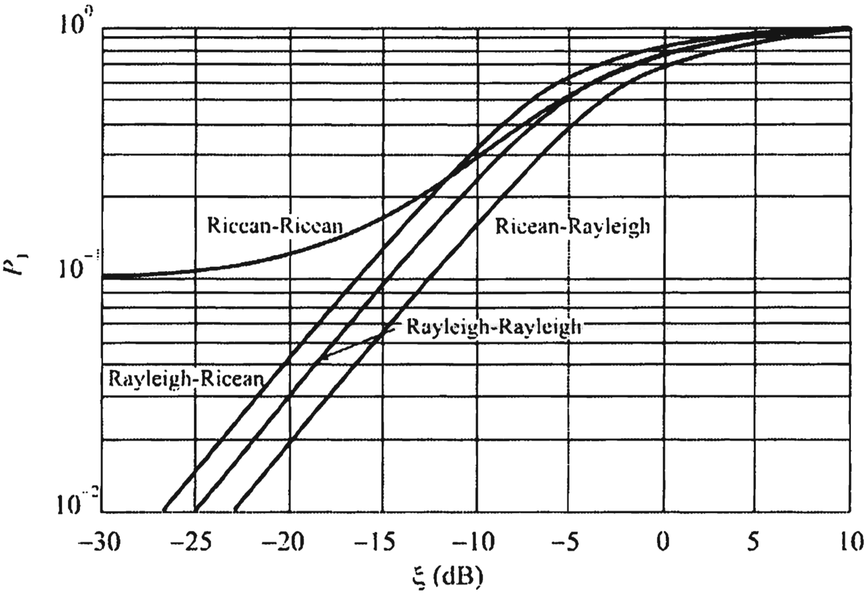


Figure 13.6 Average probability of jamming for various combinations of fading.

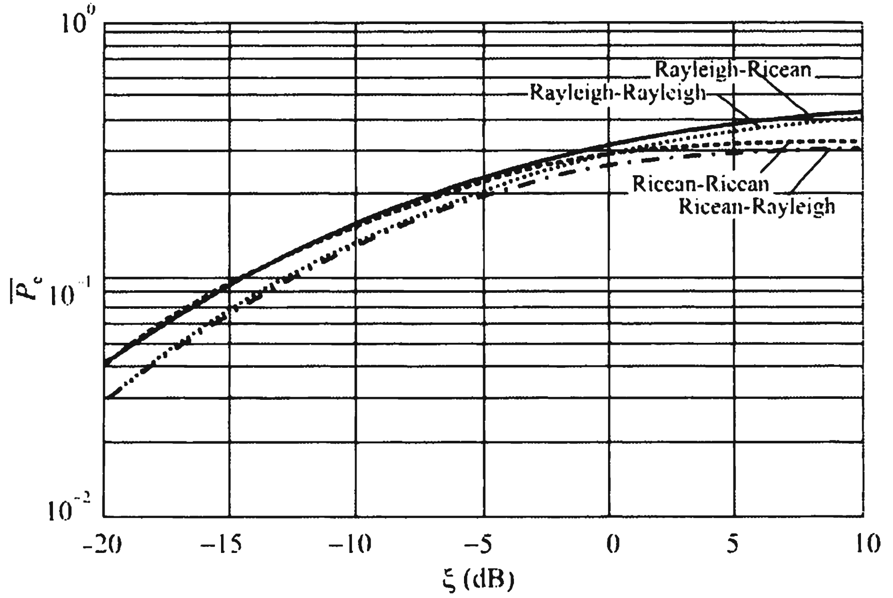


Figure 13.7 Jamming effectiveness against NCBFSK with fast fading. Ordinate is average P_e, abscissa is JSR in dB.

$$P_e(\gamma) = \frac{1}{2}\exp\left(-\frac{\gamma}{2}\right) \tag{13.19}$$

We can therefore evaluate the average probability of error by finding the average based on the pdf as

$$\bar{P}_e = \int_0^\infty f_Z(z) P_e(z) dz \tag{13.20}$$

yielding:

$$\bar{P}_e = \frac{1}{2}\int_0^\infty f_Z(z)\exp\left(-\frac{z}{2}\right)dz \tag{13.21}$$

The jamming effectiveness against BFSK is illustrated in Figure 13.7. There is not much difference in the fading model employed for most of the range of interest to this analysis. When the jamming effectiveness criteria achieves a BER of 10^{-1}, we see that at $\xi \sim -14$ to -12 dB or greater this goal is achieved. This shows that NCBFSK in reasonably fast fading channels is quite vulnerable to jamming.

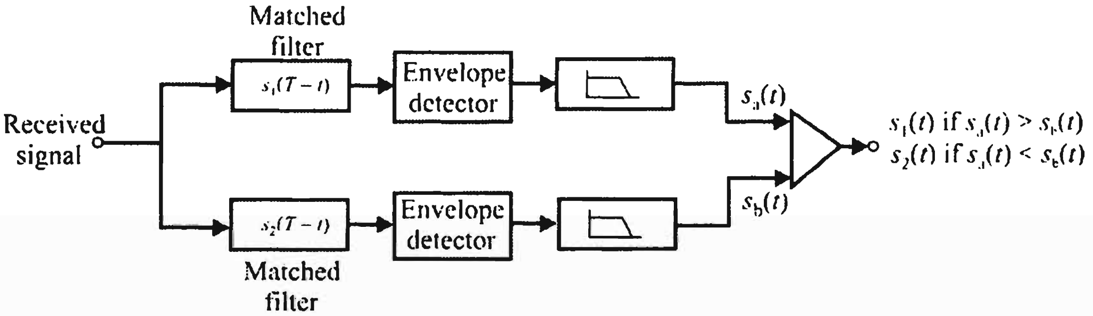


Figure 13.8 NCFSK detector. (*Source:* [7]. © IEEE 1963. Reprinted with permission.)

13.3.2 Jamming NCBFSK Signals with Narrowband Tone Jammer

A detector for narrowband NCBFSK is illustrated in Figure 13.8 [7]. The NCBFSK target signal has a bandwidth-bit duration product, BT_b, approximately equal to one (e.g., $B = 10$ kHz and $T_b = 100$ μs would be representative numbers). The jammer signal is assumed to be a random BFSK signal at the same signaling rate and with the same modulating tone frequencies. The target signal and jammer signal are assumed to be subject to independent Rayleigh fading with a rectangular spectral density of width $2B$.

An alternate detector architecture referred to as quadrature detection is illustrated in Figure 13.9. This type of detector is useful when coherent detection is not possible. A pair of modulators are driven by the LO in phase quadrature. The outputs of the I and Q multipliers are low-pass filtered to eliminate the sum terms, and applied to matched filters. These filters are matched to the mark and space of the desired transmitted signal. The output of the matched filters are cross-correlated with local replicas of the possible transmitted orthogonal signals. For the case considered here, NCBFSK, there are four possible outputs. After correlation, the outputs are squared and added to the like path in the other channel. The outputs of these two adders are evaluated, and the largest is selected as the symbol sent.

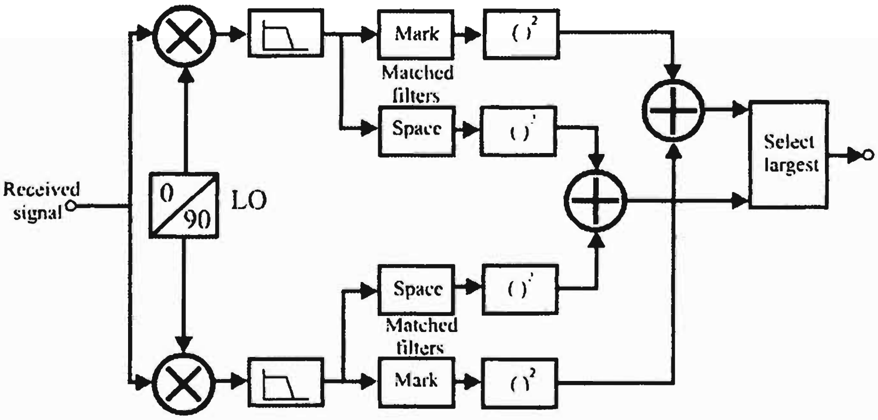


Figure 13.9 Quadrature detector for NCBFSK. (*Source:* [7]. © IEEE 1963. Reprinted with permission.)

Thermal noise is also included. For these purposes it is the accumulation of all noise sources that are not subject to fading, and normally would be dominated by the system noise as reflected in the noise figure.

The product BT_b is referred to as the fading parameter, or fading rate, and is an indication of the amount of fading present. For small BT_b, the fading is principally slow. For larger BT_b, the fading is faster.

For NCBFSK, the mean error probability is [8]

$$\bar{P}_e = \frac{1}{2}\left\{\left[1+\frac{\dfrac{N}{J}+g(BT_b)(1+1/\xi)}{\dfrac{N}{J}+\alpha(BT_b)(1+1/\xi)}\right]^{-1}+\left[1+\frac{\dfrac{N}{J}+\dfrac{g(BT_b)}{\xi}+\alpha(BT_b)}{\dfrac{N}{J}+\dfrac{\alpha(BT_b)}{\xi}+g(BT_b)}\right]^{-1}\right\} \tag{13.22}$$

The ratio N/J is the ratio of the average power of the thermal noise to the average power of the fading jamming FSK signal and ξ is the mean JSR. The functions $a(BT_b)$ and $g(BT_b)$ are defined in the following.

As the fading rate increases, the mark and space signal spectra for OOK spread out. Figure 13.10 contains spectra of an OOK carrier of frequency f_0 subject to fading for BT_b values of 0.1, 0.5, 1, 2, 3, and 4. These spectra are based on an OOK carrier which is bandpass filtered in the range ($f_0 - 1/2T_b$, $f_0 + 1/2T_b$) before transmission, corresponding to a transmitted mark or space signal in FSK transmissions. We can see that the fading signal spectrum flattens and spreads out with increasing BT_b.

Figure 13.11 shows the spectrum spreading for fast fading BFSK. The FSK receiver, illustrated in Figure 13.8, consists of a pair of contiguous ideal bandpass

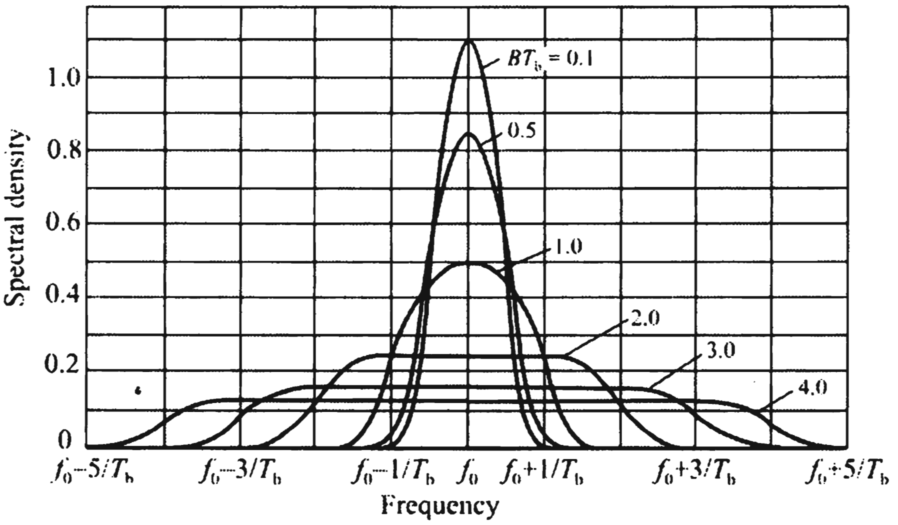


Figure 13.10 Spectral densities of Rayleigh fading OOK signals. (*Source:* [7]. © IEEE 1963. Reprinted with permission.)

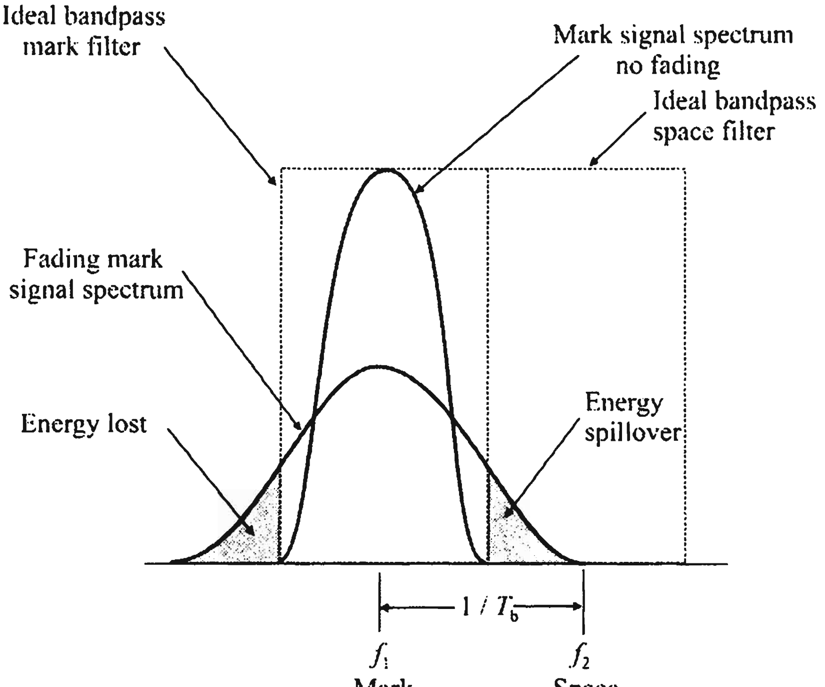

Ideal bandpass mark filter
Mark signal spectrum no fading
Ideal bandpass space filter
Fading mark signal spectrum
Energy lost
Energy spillover
1 / T_b
f_1
f_2
Mark
Space

Figure 13.11 Spectrum of mark for fast fading binary FSK signals. (*Source:* [7]. © IEEE 1963. Reprinted with permission.)

filters located in the intervals $(f_1 - 1/2T_b, f_1 + 1/2T_b)$ and $(f_2 - 1/2T_b, f_2 + 1/2T_b)$, respectively. The frequency spacing is equal to $1/T_b$. The effect of the spectrum spreading is a loss of signal energy at the output of the receiver band-pass filters. This loss function is plotted in Figure 13.12 and is denoted by the function $g(BT_b)$. Another effect is that the mark and space signal spectra now overlap. Figure 13.11 shows the typical overlap between mark and space signal spectra. The function $\alpha(BT_b)$ in Figure 13.12 corresponds to the proportion of the mark signal energy that appears in the space filter output. The third function in Figure 13.12 is the ratio $f(BT_b)$ of $a(BT_b)$ to $g(BT_b)$:

$$f(BT_b) = \frac{\alpha(BT_b)}{g(BT_b)} \tag{13.23}$$

As BT_b increases this ratio approaches unity, corresponding to complete system degradation.

When the thermal noise is negligible compared with the fading jamming, the mean error probability function in (13.22) reduces to

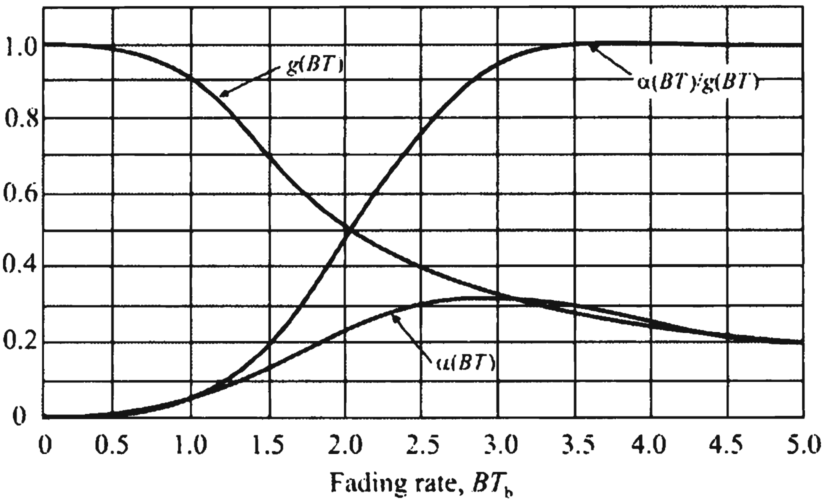


Figure 13.12 Energy loss functions due to signal spectrum spreading. $g(\cdot)$ corresponds to the loss due to the bandpass filters and $\alpha(\cdot)$ corresponds to the amount of power transferred to the opposite bandpass filter output due to the spectra spreading caused by fading. (*Source:* [7]. © IEEE 1963. Reprinted with permission.)

$$\bar{P}_e = \frac{1+(1+2/\xi)f(BT_b)}{2(1+1/\xi)[1+f(BT_b)]} \tag{13.24}$$

Equation (13.24) is plotted in Figure 13.13 for representative small values of BT_b corresponding to a narrowband FSK jamming signal. We can see that the faster the fading occurs the better we are able to successfully jam the target receivers. When $BT_b \geq 1.5$, jamming is successful with as little as –20 dB JSR.

Figures 13.14–13.18 show comparisons of the quadrature detection system (Figure 13.10) with the NCFSK detector (Figure 13.9). Figure 13.14 shows the mean error probability when $BT_b = 0$, while Figure 13.15 shows the same for when $BT_b = 1$ and when there is no separation of the filters for the FSK detector. Last, Figure 13.16 shows the mean error probability when $BT_b = 2$. Figure 13.17 shows the effect of separating the filters so there is no ISI for the FSK detector.

In general we see that the NCBFSK detectors are somewhat more resistant to noise levels (and therefore to noise jammers, too). If the space and mark filters for the NCBFSK detector were separated so that the "spillover" energy is negligible or $f(BT_b)$ were negligible, then we can see from (13.24) that the mean probability of error, when the nonfading noise is negligible compared to the fading noise (jammer signal), is.

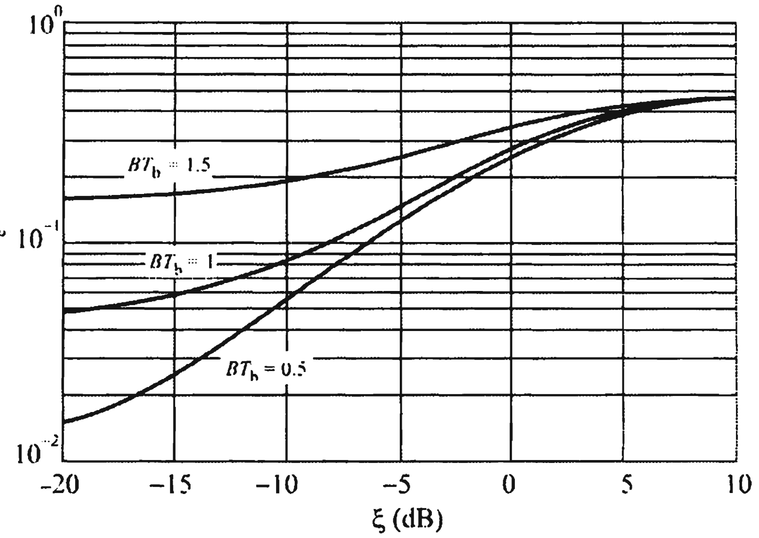


Figure 13.13 BER for NCFSK when $J >> N$.

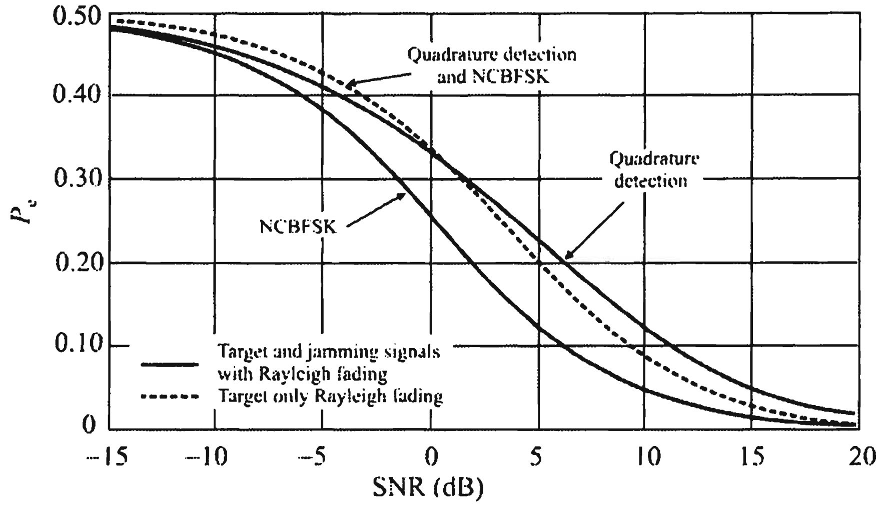


Figure 13.14 Comparison of quadrature detection and NCBFSK when $BT_b = 0$. (*Source:* [7]. © IEEE 1963. Reprinted with permission.)

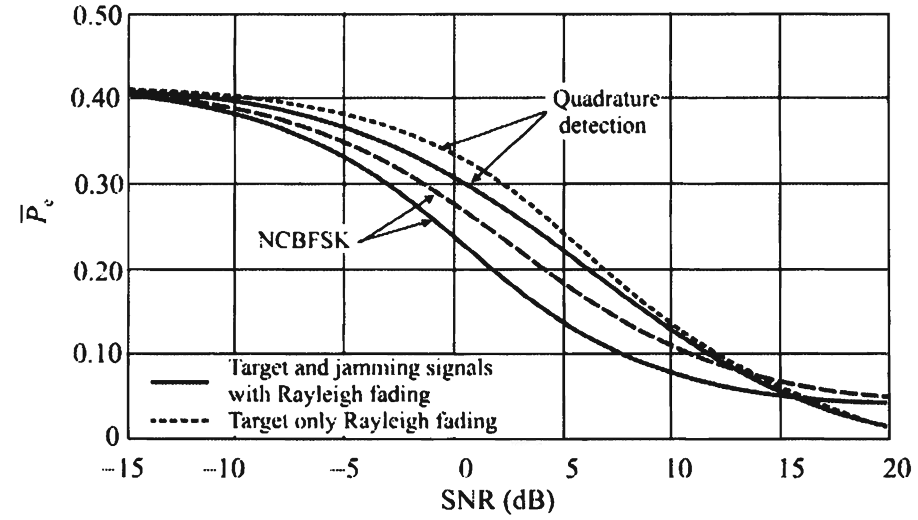


Figure 13.15 Comparison of quadrature detection and NCBFSK when BT_b = 1. (*Source*: [7]. © IEEE 1963. Reprinted with permission.)

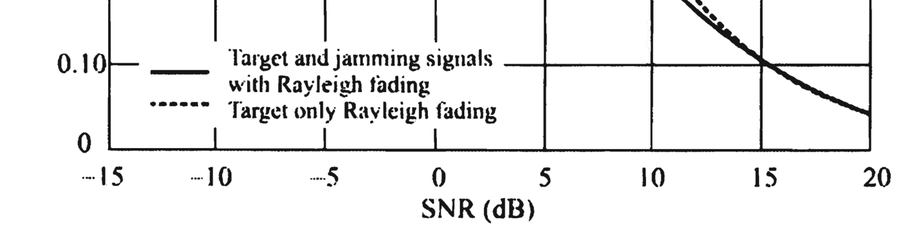


Figure 13.16 Comparison of quadrature detection and NCBFSK when BT_b = 2. (*Source:* [7]. © IEEE 1963. Reprinted with permission.)

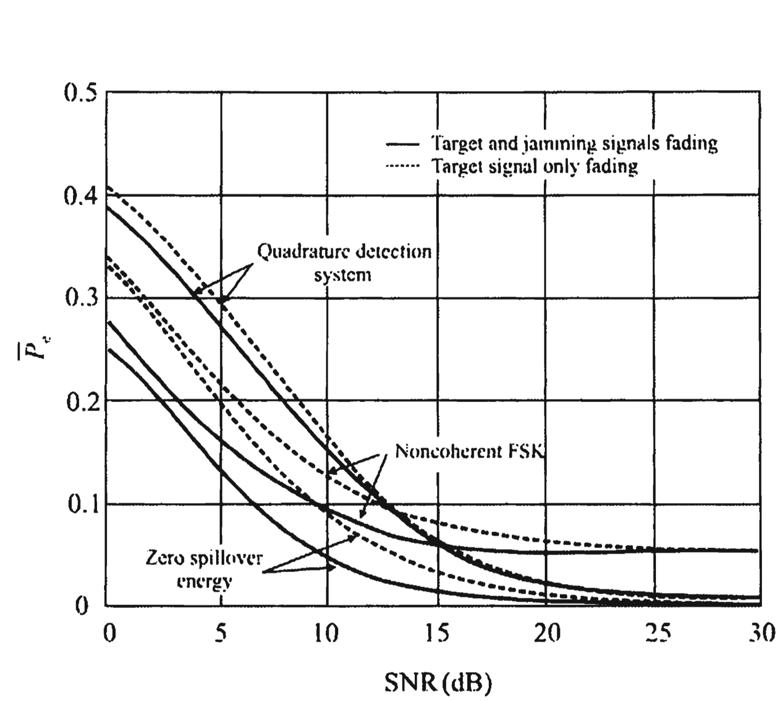


Figure 13.17 Comparison of quadrature detection and NCFSK systems as a function of SNR in dB, for $BT_b = 1$. (Source: [7]. © IEEE 1963. Reprinted with permission.

$$\bar{P}_e = \frac{1}{2\left(1 + \frac{1}{\xi}\right)} \tag{13.25}$$

We can see that the FSK detector is always superior to the quadrature detection system. Figure 13.18 shows a comparison as a function of BT_b for $\gamma = 10$ dB. These curves demonstrate the importance of low ISI for the NCBFSK system. For completeness, curves are also included for the signal-only fading case.

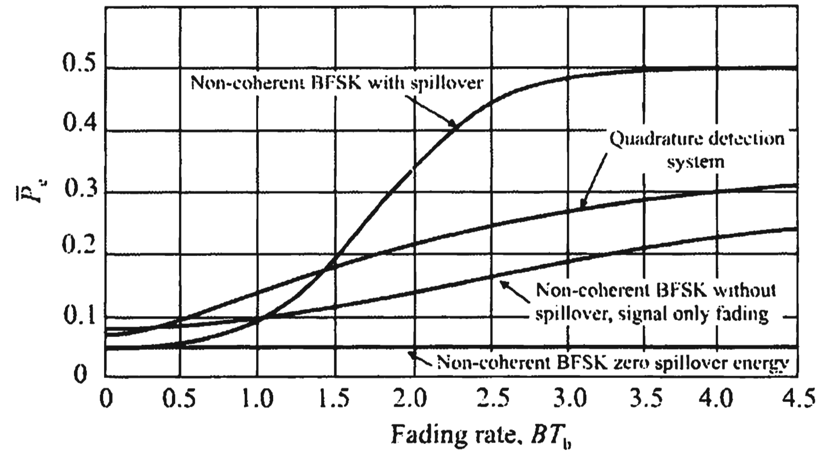


Figure 13.18 Comparison of quadrature detection and NCBFSK systems. Target and jamming signals independently fading, ideal bandpass fading spectral density, Rayleigh fading, $\gamma = 10$ dB. (*Source:* [7]. © IEEE 1963. Reprinted with permission.)

Equation (13.25) shows that $\bar{P}_\mathrm{e}$ is independent of the energy lost due to fading because both the target and jamming signals are similarly affected. Equation (13.26) shows that P_e is increased due to fading because the signal energy is decreased whereas the nonfading noise remains the same.

$$P_\mathrm{e} = \frac{1}{2+\gamma g(BT)} \tag{13.26}$$

13.3.3 Band-Limited Gaussian Noise Jammer

The first target signal we consider is wideband with a relatively high bit rate. In that case BT_b is large. The signal is detected with a quadrature detector as shown in Figure 13.9. A quadrature detector such as this is employed when the input γ is small. The jammer signal for this target consists of band-limited Gaussian noise.

For the quadrature detection system, the mean error probability is

$$\bar{P}_\mathrm{e} = \begin{cases} BT_\mathrm{b} \exp\left[\dfrac{g(BT_\mathrm{b})EBT_\mathrm{b}}{N_0+J_0}\right] E_{2BT_\mathrm{b}+1}\left[\dfrac{g(BT_\mathrm{b})EBT_\mathrm{b}}{N_0+J_0}\right], & BT_\mathrm{b} \geq 1/2 \\ \dfrac{1}{2}\exp\left[\dfrac{g(BT_\mathrm{b})E}{2N_0+J_0}\right] E_2\left[\dfrac{g(BT_\mathrm{b})E}{2N_0+J_0}\right], & BT_\mathrm{b} < 1/2 \end{cases} \tag{13.27}$$

where $E_n(x)$ is the exponential-integral function defined previously. $g(\cdot)$ was defined in Section 13.2.2. In the slow fading case, $BT_\mathrm{b} \sim 0$ and $g(\sim 0) \approx 1$, (13.27) reduces to

$$\bar{P}_\mathrm{e} = \frac{1}{2}\exp\left(\frac{E}{2N_0+J_0}\right) E_2\left(\frac{E}{2N_0+J_0}\right) \tag{13.28}$$

Equation (13.28) is plotted in Figure 13.19 for some representative values of γ.

13.4 Jamming DSSS Wideband Systems

We will discuss jamming performance in fading channels when the targets are DSSS receivers [9]. Much of this material is based on [5].

If the time difference of arrival between the signals traversing the direct path and reflected paths is small compared to the reciprocal of the spread spectrum

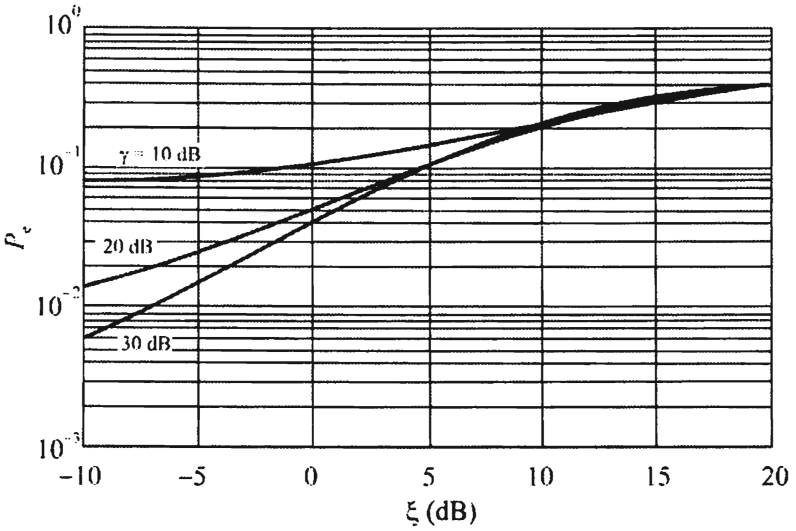


Figure 13.19 Band-limited noise jammer versus digital.

signal bandwidth, the direct path signal cannot be resolved from the indirect signals at the receiver. In this case, the results of Section 13.2.1 apply.

If, on the other hand, the time difference of arrival is much larger than the reciprocal of the signal bandwidth, the direct signal can be resolved from the indirect signals at the receiver. When the target employs DSSS, the indirect paths will be decorrelated from the direct path. The effect is to reduce the inband power of the multipath interference by the spread spectrum processing gain. The remaining interference has statistics similar to additive white Gaussian noise. If the processing gain is reasonably large (say, 30 dB or more), the interference caused by the indirect signals has little effect on the target system performance assuming that the receiver's PN generator correctly synchronizes to the direct path and not to one of the weaker indirect paths. With this assumption, at baseband (after decorrelation) the interference from other paths will be at least 30 dB below the desired signal level and will have little effect on AJ performance.

If the jammer signal is wideband (bandwidth on the order of the target signal), the interference from the indirect paths will have little effect on the direct path jamming signal so conventional AJ analysis techniques for nonfading (no multipath) signals can be employed to ascertain performance.

13.4.1 BPSK/QPSK

BPSK and QPSK are popular modulations for DSSS. Denoting the SNR by γ, the BER performance for both of these modulation types is the same and is given by

$$P_e = Q\left(\sqrt{2\gamma}\right) \tag{13.29}$$

where $Q(\cdot)$ is the Gaussian Q-function[5] and γ in this case includes the processing gain provided by the DSSS system. Thus

$$\gamma = G \cdot \text{SNR} \tag{13.30}$$

Wideband signals are subject to fast fading because of the channel coherence time limitations compared to the bit time. The above analysis assumed slow fading. At 1.25 Mbps ($T_b \sim 1$ μs), in a cell system at 10 km separation, $t \sim 0.5$ μs, which are comparable numbers, there may be frequency selective interference. In those cases when the wideband signal fading is slow fading, however, using (13.29) in conjunction with (13.20) produces

$$\bar{P}_e = \int_0^\infty f_Z(z) Q(\sqrt{2z}) dz \tag{13.31}$$

With a wideband jamming signal with jamming noise power density given by J_0, depending on the method of modulating the PSK signal on the signal waveform, a reduction in the jammer broadband noise (BBN) by the fraction κ applies. Therefore for BPSK and QPSK [6, 10]

$$P_e = Q\left(\sqrt{2\frac{E_b}{N_0 + \kappa J_0}}\right) \tag{13.32}$$

where $\kappa = 0.903$ for BPSK and 0.995 for MPSK, $M > 2$, ultimately leading to

$$P_e = Q\left(\sqrt{2\frac{1}{1/G\gamma + \kappa\frac{\xi}{G}}}\right) \tag{13.33}$$

[5] The Gaussian Q-function is given by

$$Q(x) = \frac{1}{2}\text{erfc}\left(\frac{x}{\sqrt{2}}\right)$$

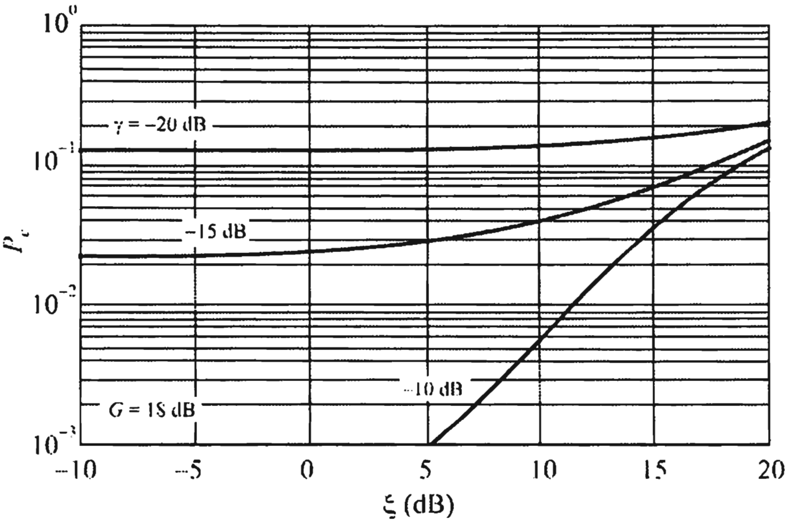


Figure 13.20 DSSS BBN jamming performance without fading.

where γ is the predecorrelation SNR ($G \cdot \gamma$ is the postdecorrelation but predetection SNR, after being low-pass filtered with bandwidth $B = R$) and ξ is the predecorrelation JSR. Without considering fading on either of the links, the BER is shown in Figure 13.20 versus ξ for G = 18 dB. Thus, when γ = −20 dB, for example, the postdecorrelation but predetection SNR is −20 + 18 = −2 dB, when it is −10 dB the predetection SNR is −10 + 18 = 8 dB, etc. For $\gamma \geq -15\text{dB}$ or so, the BER of 10^{-1} is achieved at about the point where ξ gets large enough to overcome the processing gain.

Assuming the jamming waveform is wideband Gaussian noise with bandwidth at least as wide as the target signals, and that $J_0 >> N_0$, the jamming performance is illustrated in Figure 13.21. Notice the jamming tolerance of the target network; considerably more jamming power is required for wideband targets when the jamming signal is also wideband. This is due to the spreading of the jamming signal across a considerably wider bandwidth. The decorrelation process does not collapse the jamming signal as it does for the signal with the correct PN code. For PSK, the decorrelation process decreases the jamming signal level by the factor κ as mentioned. So when fading considerations are included it is significantly more difficult to effectively jam DSSS targets.

13.4.2 Comments

The analysis of jamming performance against narrowband and wideband targets when both links experience Ricean fading was presented in this section. This analysis used a pdf corresponding to the ratio of two Ricean r.v.s. The fading

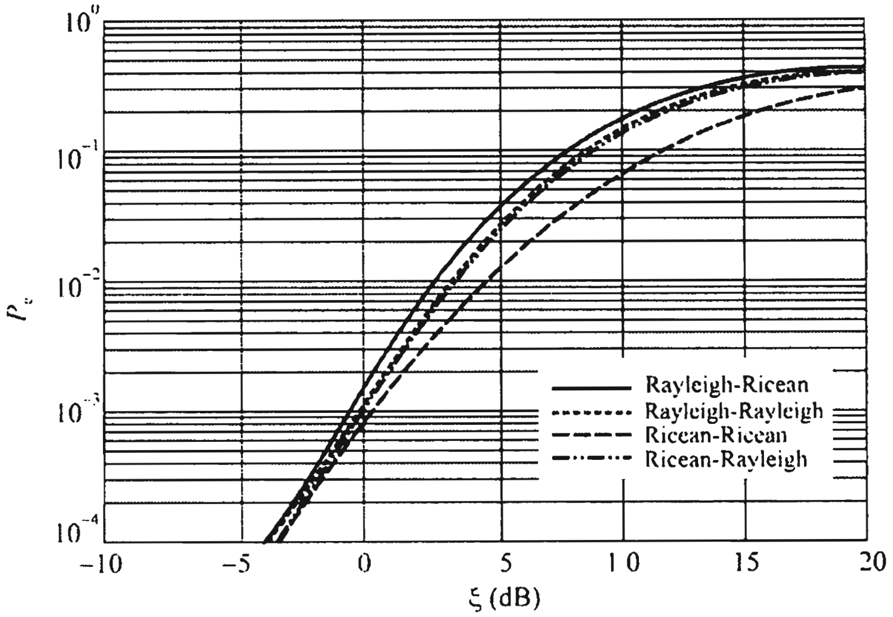


Figure 13.21 Example of wideband jamming of QPSK signal. $C_s = C_J = 1$ and $\sigma_J = 1$. PG = 13 dB.

invokes a requirement to provide more AJ margin than when fading is not considered. Typically the required AJ processing gain increases by 10–20 dB, depending on the required communications availability and the type of fading. The two principal restrictions for these results to apply are that the fade rate must be slow compared to the data rate and that the signal bandwidth must be small enough that the fading can be considered flat (non-frequency selective).

13.5 Jamming Performance in Nakagami Fading

This section presents an analysis of jamming performance when the channel between the target transmitter and receiver as well as the channel between the jammer and the target receiver exhibit Nakagami-*m* fading characteristics [11]. This section is based on an adaptation of the developments in Wojner [12] and Al-Hussaini [13], where the measure of effectiveness (MOE) is jamming success versus communication success.

13.5.1 Slow Fading

Assume that both the signal and jammer exhibit non-frequency-selective fading and the pdfs of their envelopes are adequately represented by the Nakagami distributions. The pdf of the power of the signal is then given by (2.66) repeated here for convenience as [13]

$$f_W(w) = \left(\frac{m}{\bar{w}}\right)^m \frac{w^{m-1}}{\Gamma(m)} \exp\left(-\frac{mw}{\bar{w}}\right), \qquad w \geq 0, m \geq \frac{1}{2} \tag{13.34}$$

Let γ_J denote the SJR so that $\xi = 1/\gamma_J$ and both S and J have the distribution defined by (13.34), that is $w \in \{S, J\}$, but, in general, with different fading parameters, m_S and m_J, respectively. $\bar{S}$ denotes the average target signal power at the target receiver while $\bar{J}$ denotes the same for the jammer. Assuming that both the target signal and the jamming signal are large enough that noise can be ignored then the probability of jamming the target is given by

$$P_J \triangleq \Pr\{\bar{S} \leq c\bar{J}\} \tag{13.35}$$

for some suitable c. Then

$$P_J = \Pr\{\gamma_J \leq c\} = F_{\Gamma_J}(c) \tag{13.36}$$

where $F_{\Gamma_J}(c)$ is the cumulative distribution function of γ_J defined as

$$F_{\Gamma_J}(c) = \int_{-\infty}^{c} p_{\Gamma_J}(x)dx \tag{13.37}$$

when $p_{\Gamma_J}(\gamma_J)$ is the pdf of γ_J. Since it is the only case of interest, we assume that $\bar{S} \geq 0$ and $\bar{J} \geq 0$ so that the lower limit in (13.37) can be set to zero,[6] so that

$$F_{\Gamma_J}(c) = \int_{0}^{c} p_{\Gamma_J}(\gamma_J)d\gamma_J \tag{13.38}$$

The pdf of γ_J is thus given by [13]

[6] This assumption does not imply that $\gamma \geq 0$ or that $\xi \geq 0$.

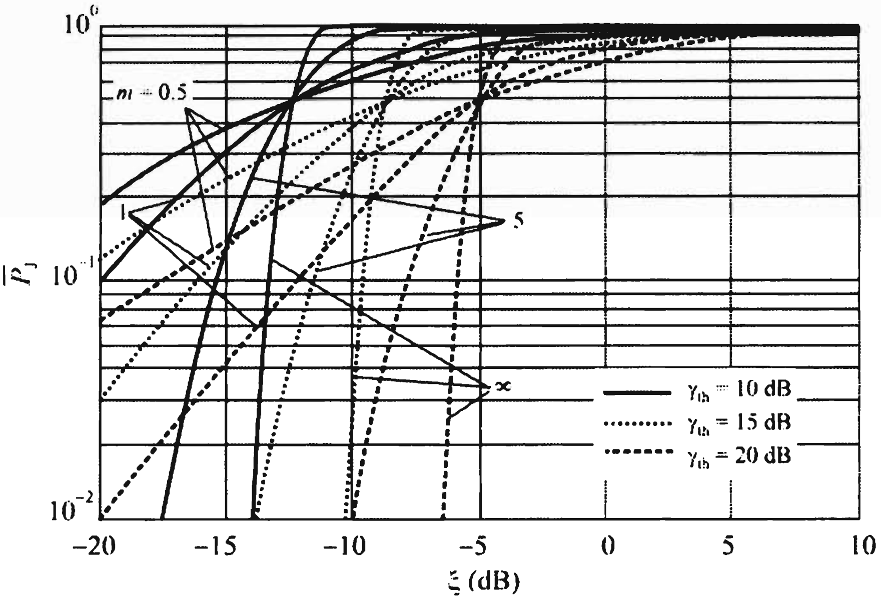


Figure 13.22 Nakagami slow fading channel for some typical threshold values. ξ is the JSR in dB $m_S = m_J = m$.

$$p_{\Gamma_J}(\gamma_J) = \frac{h^{m_S}\gamma_J^{m_S-1}}{B(m_S,m_J)(1+h\gamma_J)^{m_S+m_J}} \tag{13.39}$$

where

$$h = \frac{m_S}{m_J}\xi$$

and

$$B(m_S,m_J) = \frac{\Gamma(m_S)\Gamma(m_J)}{\Gamma(m_S+m_J)} \tag{13.40}$$

$B(m_S, m_J)$ is the beta function [14]. The average probability of jamming success is given by the incomplete beta function defined as [15]

$$\bar{P}_J(m_S,m_J) = \frac{1}{B(m_S,m_J)}\int_0^k u^{m_S-1}(1-u)^{m_J-1}du \tag{13.41}$$

where

$$k = \frac{\gamma_{\text{th}} m_{\text{S}}}{\gamma_{\text{th}} m_{\text{S}} + {}^{m_{\text{J}}}\!/\!_{\xi}}$$

γ_{th} is a threshold below which communication is ineffective, or, in other words, jamming is successful.

Equation (13.41) can be solved numerically to find the probability of successful jamming as m_{S}, m_{J}, and γ_{th} are varied. The results are illustrated in Figure 13.22. With Nakagami fading present, the jamming performance is improved considerably compared to when it is absent. An intuitive explanation for this is that for successful digital communications most of the bits must get passed error free. For a BER of 10^{-3}, 99.9% of the bits must be successfully communicated. On the other hand, it can be shown [16] that achieving a BER of 10^{-1} by employing jamming can preclude communication in such systems, even when popular forms of FEC coding are applied to the communication signal. Therefore the jammer can be effective much less of the time than the communication system and still be successful. Since fading is detrimental to communications, its effect on the target link is more dramatic than on the jamming link.

13.5.2 Faster Fading

In this case it is assumed that the fading is slow compared to the data rate but fast compared to the message duration. In this case the average probability of error is given by

$$\bar{P}_{\text{e}} = \frac{1}{2}\int_{0}^{\infty} p_Z(z) e^{-az} dz \qquad (13.42)$$

where $p_Z(z)$ is given by (13.39) and a is defined by

$$a = \begin{cases} 1/2 & \text{for NCFSK} \\ 1 & \text{for DPSK} \end{cases} \qquad (13.43)$$

Numerical evaluation of (13.42) yields the results shown in Figure 13.23 for some values of $m_{\text{S}} = m_{\text{J}} = m$ as ξ was varied. Again, as in the case of slow fading, the Nakagami fading had a significant effect on the ability to jam the target link. When the fading was severe, say $m \le 1$, the rate at which jamming is effective improves linearly as opposed to exponentially when the fading was light or

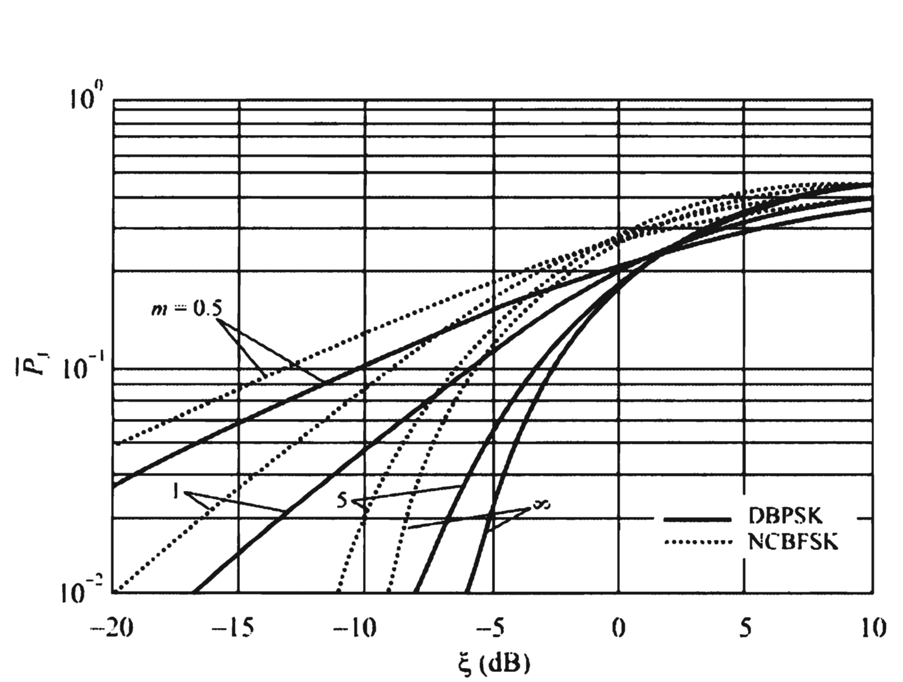


Figure 13.23 Average probability of jamming with moderate rate Nakagami fading: dotted lines, NCFSK; solid lines, DPSK ($m_S = m_J$).

nonexistent. This is a significant advantage for the jammer. Also note that in general NCBFSK is more vulnerable to jamming than DBPSK.

13.6 Jamming FHSS Wideband Systems

Simon et al. determined that the optimal jammer strategy for FHSS target signals is wideband jamming when the target experiences Rayleigh fading and the jamming signal did not fade [16]. As discussed above, Wojnar [12] evaluated Nakagami-*m* faded signals with unfaded interference. This section considers wideband and optimized partial-band jamming when either the signal or the signal and jammer experience Nakagami-*m* fading (including Rayleigh fading as a special case). Simple BER expressions are found using a large SNR approximation. This section follows [17] fairly closely.

13.6.1 Noise Jamming

The NCBFSK BER in AWGN is

$$P_{\mathrm{e,AWGN}} = \frac{1}{2} e^{-\gamma_n/2} \tag{13.44}$$

where $\gamma_n = E_b / N_0$ is the SNR, defined as the ratio of bit energy to noise power density. This parameter can be converted to a power ratio, denoted by γ, by multiplying by the ratio of the bit rate to the noise bandwidth; the latter is typically assumed to be the channel bandwidth. The signal is assumed to frequency hop across a W_{ss} Hz bandwidth. Each bit is assumed to be transmitted at a single frequency, which is frequently defined as *fast frequency hopping* (FFH), although it is generally accepted as the rate that divides *slow frequency hopping* (SFH) from FFH. If multiple bits are transmitted at one frequency (SFH), it is assumed that interleavers randomize their order in the bit stream, which is normally the case. Randomizing the bits moves them around in the bit stream relative to one another and improves the BER performance of the communication link.

The jammer considered here is a wideband jammer that emits a Gaussian noise-like signal with power distributed uniformly across the W_{SS} Hz bandwidth (BBN jamming). We will assume that the noise jammer signal, when it is present, is sufficiently more powerful than the noise so that the latter can be ignored. Denote the jammer power by J, then the jammer power density $J_0 = J / W_{SS}$. The *signal-to-jam ratio* (SJR) is defined $\gamma_J = E_b / J_0 = GS / J$, where S is signal power, $1/T$ is the data rate, and $G = W_{SS}T$ is the spread spectrum processing gain. We use ξ to denote the jam-to-signal ratio, $\xi = 1/\gamma_J$. If $J_0 >> N_0$, the wideband jammer BER is approximated by (13.44) using γ_J in place of γ. A *partial-band noise* (PBN) jammer that concentrates its power in a βW_{SS} Hz sub-band of W_{SS}, $0 < \beta \le 1$, will jam the signal with probability β. When the signal is jammed, effective jammer power density is J_0 / β. β is called the *fractional jammer bandwidth*. If the jammed BER grows faster than the fractional bandwidth falls, smaller values of β will result in higher BER. For a more complete discussion of PBN jamming see [6].

When γ is much larger than γ_J, the error probability in the unjammed bandwidth can be ignored, and the overall error probability is approximated by

$$P_{\text{e,PBN}} = \frac{\beta}{2} e^{-\beta\gamma_J/2} \tag{13.45}$$

Maximizing (13.45) in terms of β, the optimal fractional bandwidth is $\beta^* = \min(1, 2/\gamma_J)$, yielding the effective BER [6, 17]

$$P^*_{\text{e,PBN}} = \begin{cases} \frac{1}{2} e^{-1/2\xi}, & \beta = 1 \\ e^{-1}\xi, & \beta < 1 \end{cases} \tag{13.46}$$

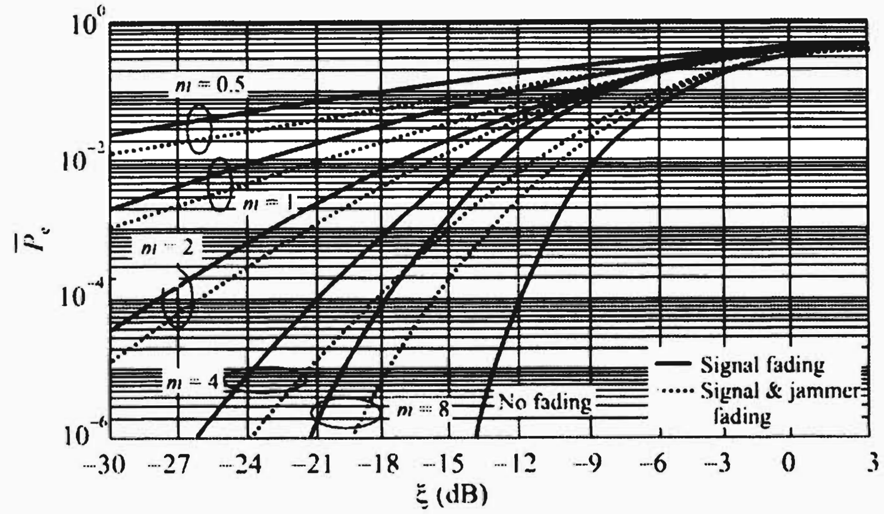


Figure. 13.24 Approximate BER for a faded signal subject to wideband jamming, for different values of signal fading parameter *m*. Jamming is faded or unfaded, with identical signal and jamming fade statistics when present. (*Source:* [17]. © IEEE 2003. Reprinted with permission.)

13.6.2 Signal-Only Fading

When only the signal is subject to fading, and the dominant interference is either AWGN or an unfaded jammer, γ or γ_J become $\gamma = \alpha_\text{S}\overline{\gamma}$, where α_S is a unity-mean signal fading term with the Nakagami-*m* distribution, and $\overline{\gamma}$ is the mean γ or γ_J. The fading loss varies randomly as the target signal hops, and with jamming the BER is the average error rate over fading [6]

$$\begin{aligned} P_\text{fc} &= \frac{m^m}{2\Gamma(m)} \int_0^\infty \alpha_\text{S}^{m-1} \exp\left[-\left(m + \frac{1}{2\xi}\right)\alpha_\text{S}\right] d\alpha_\text{S} \\ &= \frac{1}{2}\left(\frac{m}{m + {}^{1}\!/_{2\xi}}\right)^m \end{aligned} \tag{13.47}$$

When $m = 1$, this expression agrees with results for Rayleigh faded signals in jamming [12, 17]. Figure 13.24 shows BER curves for a fading signal in wideband jamming, using (13.47). Curves are also shown for signal and jammer fading, using results from the next section. For small values of *m*, fading is severe, and the BER increases slowly as ξ increases. As *m* grows, performance approaches the unfaded signal curve.

It was shown in [17] that the optimal strategy for jamming a Rayleigh faded target signal is wideband jamming. The same is not necessarily true for Nakagami-

m fading with $m > 1$. Again assuming $J_0 >> N_0$, the BER in partial-band jamming with fractional bandwidth β is given by

$$P_{\mathrm{fc}}(\beta) = \frac{\beta}{2}\left(\frac{m}{m + \beta/2\xi}\right)^{m} \tag{13.48}$$

The optimal jammer bandwidth based on (13.48) is

$$\beta^{*}_{\mathrm{fc}} = \begin{cases} \min\left(1, 2\xi\dfrac{m}{m-1}\right), & m > 1 \\ 1, & m \le 1 \end{cases} \tag{13.49}$$

When $m \le 1$, the resulting optimal strategy is wideband jamming, as expected. When $m > 1$, the optimal strategy will be wideband jamming when ξ is large. As ξ decreases, the optimal jammer bandwidth narrows. In the small ξ region, the optimal partial-band jamming BER becomes

$$P^{*}_{\mathrm{fc}}\left(\xi < \frac{m-1}{2m}\right) = \xi\left(\frac{m-1}{m}\right)^{m-1} \tag{13.50}$$

Figure 13.25 is a plot of the optimized jamming fractional bandwidth β assuming γ is large, parameterized with m_S. Curves are also shown using results from the following section, for the case where both target signal and jammer signal are subjected to fading. It is assumed that the signal and jammer fading parameters are equal, $m_J = m_S$. When $m_S \le 1$, the optimal jamming technique is wideband for all ξ. When $m_S > 1$, the optimal bandwidth increases as the inverse of ξ, once ξ exceeds $(m - 1)/2m$. Figure 13.26 is a plot of the resulting BER curves. The BER rises above 10^{-1} in the range of $\xi \sim -12$ to -6 dB, depending on the fading parameter.

13.6.3 Both Target Signal and Jammer Signal Fading

We would expect that if the target signal fades so would the jammer signal, especially since they are typically relatively close together. In this section we examine how jammer signal fading affects the jamming performance. The two signals would likely experience fading with different shape parameters, m_S for the

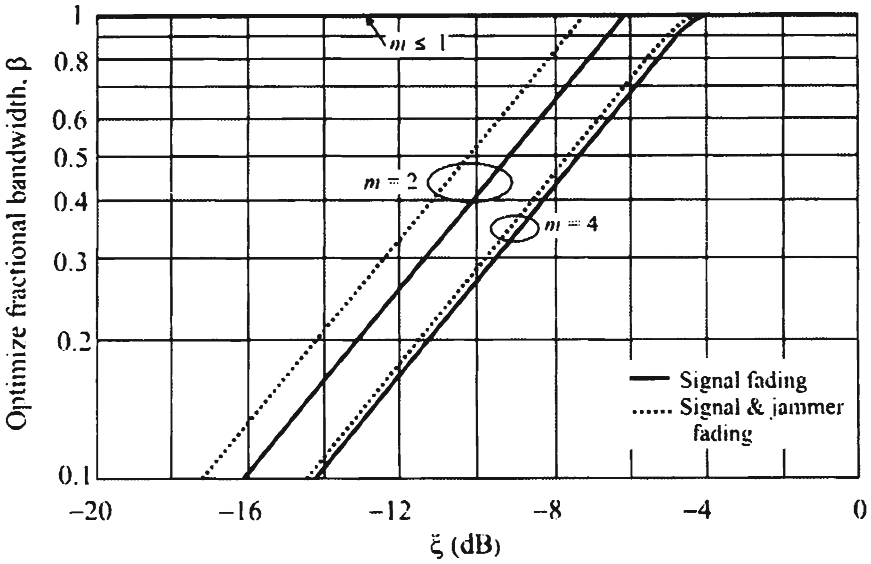


Figure 13.25 Optimal fractional bandwidth, β, for a fading signal subject to partial-band jamming, for different values of signal fading parameter *m*. Jamming is fading or nonfading, so that $m = m_S = m_J$ when the jammer is fading. (*Source:* [17]. © IEEE 2003. Reprinted with permission.)

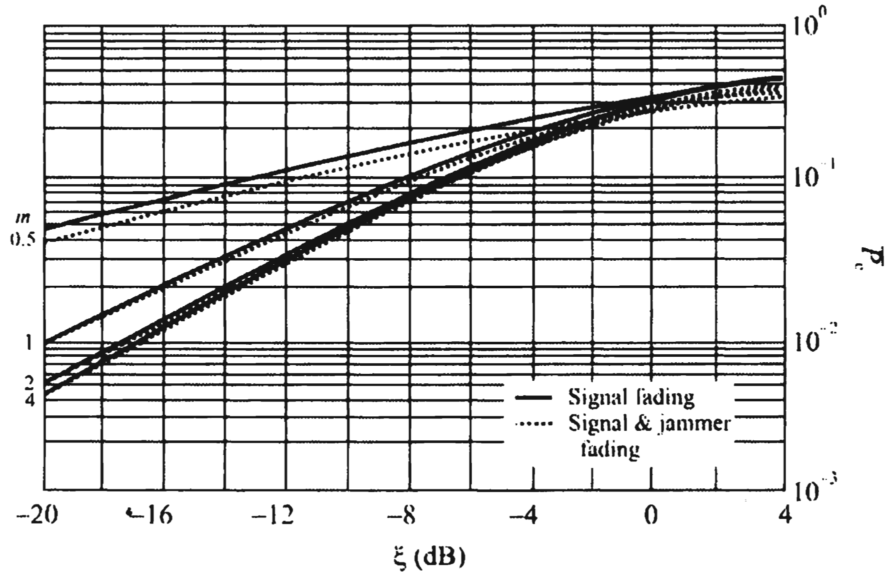


Figure 13.26 BER for a faded signal subject to optimized partial-band jamming for different values of signal fading parameter *m*. Jamming is fading or nonfading, with identical signal and jamming fade statistics when present. We can see that including the effects of the jamming signal fading has little effect on the BER (about 1 dB at lower ξ). (*Source:* [17]. © IEEE 2003. Reprinted with permission.)

signal, and m_J for the jammer, because normally the target transmitter, receiver, and jammer transmitter would not form a linear line, a scenario that could lead to the same channel for the target signal and jammer signal. The propagation ranges are likely different between the receiver-transmitter and receiver-jammer links as well. Conversely, a small distributed jammer [18] may be more severely faded than the target link (we will discuss distributed jammer systems in Chapter 15).

Both signal and jammer power are now multiplied by independent, unit-mean, fading factors, α_S and α_J, and the SJR is $\gamma_J = \rho\bar{\gamma}_J$ where $\rho = \alpha_S / \alpha_J$. As the ratio of two independent, chi-squared r.v.s, ρ is also an r.v. characterized by the F distribution [19], with pdf

$$p(\rho) = \frac{\left({}^{m_S}\!/\!_{m_J} \right)^{m_S} \rho^{m_S - 1}}{B(m_S, m_J)\left(1 + \rho^{m_S}\!/\!_{m_J} \right)} \tag{13.51}$$

where $B(m_S, m_J)$ is the beta function given by (13.40).

In wideband jamming, when γ is large then the BER is (11)

$$P_{ff} = \frac{1}{2}\int_0^\infty \exp\left[-\rho \frac{1}{2\xi} \right] \frac{\left({}^{m_S}\!/\!_{m_J} \right)^{m_S} \rho^{m_S - 1}}{B(m_S, m_J)\left(1 + \rho^{m_S}\!/\!_{m_J} \right)} d\rho \tag{13.52}$$

Using the confluent hypergeometric function $U(a, b, z)$[7] [20] this becomes,

$$P_{ff} = \frac{\Gamma(m_S + m_J)}{2\Gamma(m_J)} U\left(m_S, 1 - m_J, \frac{m_J}{m_S} \frac{1}{2\xi} \right) \tag{13.53}$$

In the special case of Rayleigh fading (no direct component) for both signal and jamming, $m_S = m_J = 1$, this BER becomes

$$P_{ff}(m = 1) = \frac{1}{2} e^{1/2\xi} E_2\left(\frac{1}{2\xi} \right) \tag{13.54}$$

[7] The integral definition of the confluent hypergeometric function is given by

$$U(a, b, z) \triangleq {}_1F_1(a; b; z) = \frac{\Gamma(b)}{\Gamma(a - b)\Gamma(a)} \int_0^1 e^{2x} x^{a-1} (1 - x)^{b-a-1} dx$$

where $E_2(\cdot)$ is the exponential integral function[8] with $n = 2$ [21]. Wideband jamming results using (13.53) are shown in Figure 13.24, along with results for unfaded jamming. For convenience $m_S = m_J$ in this figure. For small m (severe fading environments), jammer signal fading decreases jamming performance by more than 4 dB. For large m, the degradation is approximately 2 dB.

Ignoring errors that occur in the unjammed portion of the hopping bandwidth, (13.53) can be modified to reflect partial band jamming

$$P_{\text{ff}}(\beta) = \beta \frac{\Gamma(m_S + m_J)}{2\Gamma(m_J)} U\left(m_S, 1 - m_J, \frac{m_J}{m_S} \frac{\beta}{2\xi} \right) \tag{13.55}$$

The effectiveness of partial-band jamming is maximized when the derivative of this equation with respect to β equals 0:

$$\frac{\Gamma(m_S + m_J)}{2\Gamma(m_J)} \left[U(m_S, 1 - m_J, z) - m_S z U(1 + m_S, 2 - m_J, z) \right] = 0 \tag{13.56}$$

where $z = m_J\beta / 2\xi m_S$. This equation must be solved numerically to find the optimal value, z^*. The optimal jamming strategy is

$$\beta^* = \min(1, 2m_S \xi z^* / m_J) \tag{13.57}$$

Using [25], (13.56) can be rewritten as

$$m_S(m_S + m_J)U(1 + m_S, 2 - m_J, z) = (m_S - 1)U(m_S, 1 - m_J, z) \tag{13.58}$$

The function $U(a, b, z)$ is positive for all sets of input arguments $a > 0$ and $z > 0$, showing that there is no finite z^* that meets the optimality condition when $m_S \leq 1$, and therefore optimal jamming is wideband. At $z = 0$ the left-hand expression in (13.58) is larger than the right-hand expression. As $z \to \infty$ the reverse is true, showing there is a solution for z^* when $m_S > 1$. Whether the optimal jamming strategy will be wideband or not depends on the product $z^* \xi$.

[8] The exponential integral is given by

$$E_n(x) = \int_1^{\infty} \frac{e^{-xt}}{t^n} dt, \qquad x > 0, n = 0, 1, \ldots$$

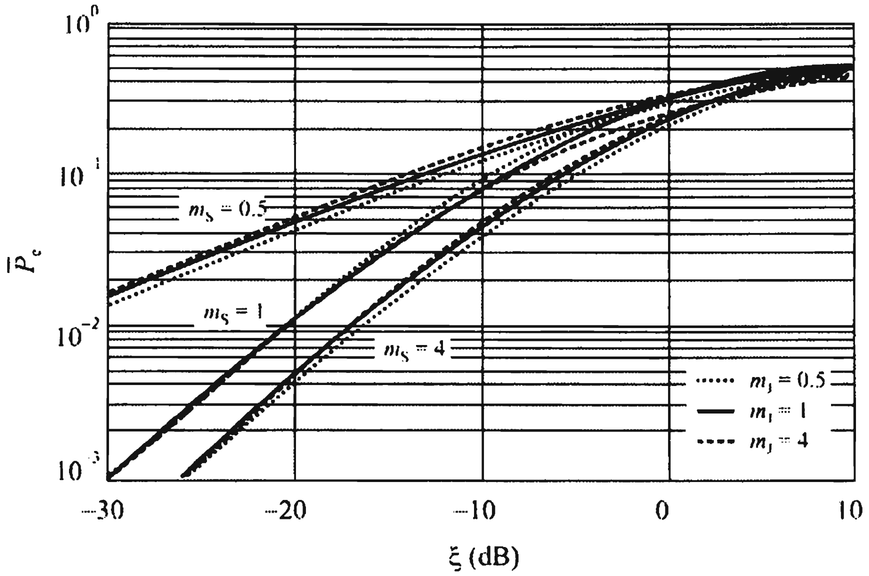


Figure 13.27 Faded partial-band jamming for different fading parameter values. BER for a fading signal subject to optimized jamming bandwidth. (*Source:* [17]. © IEEE 2003. Reprinted with permission.)

Figure 13.27 is a plot showing the BER for a faded signal, with and without jammer fading. The signal and jammer fading parameters are varied independently over the range 0.5 to 4. At high ξ, the jammer fading parameter dominates performance, but for ξ small enough so that the BER is below 10^{-1}, variations with m_J are small, and the signal fading parameter, m_S, dominates performance. As shown in Figure 13.26, there is little difference in BER performance between the faded and unfaded jammer cases in optimized partial-band jamming, when $m_S \geq 1$.

13.6.4 Remarks

Jamming performance has been evaluated for FH-NCBFSK target signals in wideband jamming, and optimized partial-band jamming, when the target signal and the jammer are subject to Nakagami-*m* fading. The optimal jammer strategy was found to be determined largely by the signal fading parameter. When there is no jammer fading, and $m_S \leq 1$, wideband jamming is optimal. Otherwise the optimal technique transitions to partial-band jamming when ξ falls below $(m_S - 1)/2m_S$ after which the optimal bandwidth falls off linearly with ξ. BER also falls off linearly with ξ in that region.

Jammer fading was found to have a noticeable effect on BER performance only in the wideband jamming case, when jammer fading decreases jammer performance by an effective power gain of 2–4 dB. When $m_S > 1$, and jamming strategy is optimized, the optimal jammer fading strategy shows some variation

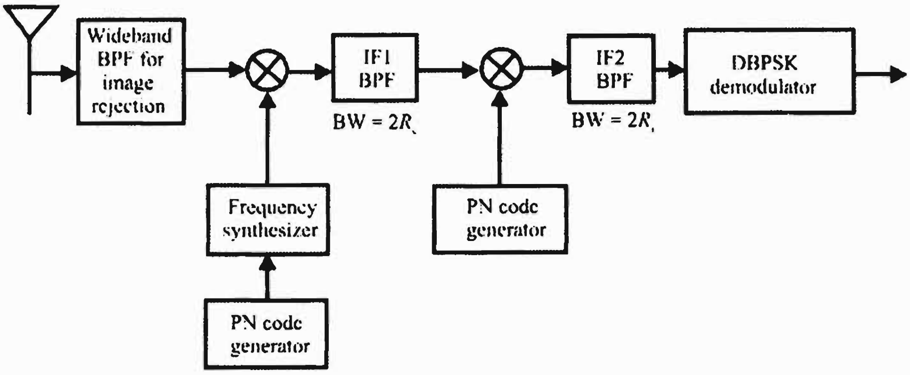


Figure 13.28 Receiver for hybrid LPI signals. (*Source:* [23]. © IEEE 1991. Reprinted with permission.)

with the jammer fading parameter, but jammer fading has very little effect on the value of the BER when the bandwidth is optimized.

13.7 Jamming Hybrid Wideband Systems

The major benefits of FHSS LPI communication systems is their difficulty in interception while maintaining a degree of simplicity in construction. The major benefits of DSSS LPI communication systems are their inherent secrecy—they are difficult to detect, however, they are complex to build. In this section we will examine a third type of LPI communication system that combines the advantages of each of FHSS and DSSS. It is called a hybrid system and incorporates features of each of the other systems [22].

We will evaluate the jamming performance by describing the average BER of the system in a jamming channel that experiences Rayleigh fading with lognormal shadowing [8]. This BER is compared with the BER with just AWGN. We will assume that DBPSK modulation is used on the target signal. (Some form of PSK is used for most, if not all, DSSS systems.)

Our receiver model is shown in Figure 13.28. The received signal is filtered for image rejection, de-hopped and filtered at the DS spread bandwidth ($2R_c$), and then DS de-spread. Since the jammer signal has been spread to a bandwidth of $2R_c$ (much greater than the data rate, R_i) the gain of the spread spectrum signaling is achieved by narrowband filtering the de-spread signal at the information bandwidth ($2R_i$).

The input to the binary DPSK demodulator during the ith signaling interval is denoted by $r_i(t)$ and is composed of three distinct components: $s_i(t)$ the message signal, $j_i(t)$ the DS spread jamming signal, and n_i(t) a narrowband Gaussian random process. Hence the received signal can be expressed as

$$r_{\text{i}}(t) = s_{\text{i}}(t) + j_{\text{i}}(t) + n_{\text{i}}(t) \tag{13.59}$$

with the jamming signal present and

$$r_{\text{i}}(t) = s_{\text{i}}(t) + n_{\text{i}}(t) \tag{13.60}$$

when it isn't.

The DBPSK demodulator compares two adjacent signaling intervals, the $(i - 1)$th and the ith, to decide what the symbol is in the latter interval. It is assumed that the IF section of the receiver does not degrade the signal. We further assume that the FH and DS PN generators are perfectly synchronized with those at the target transmitter.

The DBPSK signal in the ith signaling interval is

$$s_i(t) = \sqrt{2S}\cos(\omega_{\text{c}}t + \theta_i) \tag{13.61}$$

where θ_i can take on either of the two phases 0 and π radians, ω_{c} is the frequency of the DBPSK modulated carrier signal and for our purposes can be considered the center IF frequency, and S is the average power in the signal.

The received signal envelope $\sqrt{2S}$ fluctuates rapidly due to multipath propagation in the channel and interference from other signals. In most urban settings and some suburban settings the fluctuations approximately obey a Rayleigh pdf. In the analysis of such communication systems, it is common to normalize the signal envelope to the local mean signal level over a distance of about 50m because local means typically fluctuate by as much as 6–12 dB due to log-normal shadowing [23].

The output of the IF chain when the input is AWGN may be expressed in quadrature components as

$$n_i(t) = n_{\text{I}i}(t)\cos(\omega_{\text{c}}t) - n_{\text{Q}i}(t)\sin(\omega_{\text{c}}t) \tag{13.62}$$

where $n_{\text{I}i}(t)$ and $n_{\text{Q}i}(\text{t})$ are zero-mean, independent Gaussian random processes with 2-sided power spectral densities $N_0/2$ located in the ith signaling interval. We assume that

$$\mathcal{E}\{n_i(t)n_k(t)\} = 0, \qquad i \neq k \tag{13.63}$$

For specificity, the jammer signal is assumed to be a wideband barrage waveform. Its effectiveness against hybrid SS signals is discussed in this section.

For our purposes, the barrage waveform is assumed to have constant power spectral density, J_0, over the IF bandwidth, $2R_c$. We assume that a frequency channel is either totally jammed or not jammed at all yielding the jamming power density per channel as J_0/K where K is the number of jammed FH channels.

There are four cases that must be considered to evaluate the jamming effectiveness against DBPSK signals. There are two symbol intervals, the $(i - 1)$st one and the ith one. Each of these may or may not be affected by the jamming waveform depending on whether the channel was jammed during that interval.

The DS process at the receiver will spread the jamming signal and reduce its spectral density at the center by a scale factor. Whereas a narrowband interference signal, or a narrowband jammer, at the input to the DS decorrelation process will have its power distributed over $2R_c$, and the power level is reduced by the DS processing gain that is not true for the band-limited noise interference being considered here. The power density is reduced but not nearly by the level of the processing gain. This factor has been determined to be [24]

$$S_n = 0.9028 \tag{13.64}$$

Therefore a jamming signal with a power spectral density of $J_0/2$ before the decorrelation process has a spectral height of

$$S_n \frac{J_0}{2} \tag{13.65}$$

at the center of the second IF filter shown in Figure 13.28. Even though the decorrelation process forces the jamming spectral density to no longer have a constant amplitude, because the filtering in the second IF section is narrow, we can assume it to be so at the level specified by (13.65).

For notational convenience we will denote by J_{i-1} when the $(i - 1)$st time interval is jammed and J_i when the ith time interval is jammed. Likewise we will denote the opposite conditions by $\bar{J}_{i-1}$ and $\bar{J}_i$, respectively.

Because of the power reduction of the barrage jamming signal by the DS decorrelation process, the effective jamming power density per jammed channel is given by $S_n J_0 / K$. Furthermore, the narrowband filtering at IF2 after the DS decorrelation reduces the jamming power density even further resulting in an effective jamming power density of

$$J_e = \frac{S_n J_0 R_i}{K R_c} \tag{13.66}$$

The conditional error probabilities conditioned on whether the jammer is present or not, for DBPSK in AWGN with 2-sided spectral density of $N_0/2$, are given by

$$\Pr\{\text{error}|\bar{J}_{i-1},\bar{J}_i\} = \frac{1}{2}\exp\left[-\frac{S}{N_0}\right] \tag{13.67}$$

when neither time slot is jammed,

$$\Pr\{\text{error}|J_{i-1},J_i\} = \frac{1}{2}\exp\left[-\frac{S}{N_0+N_e}\right] \tag{13.68}$$

when both time intervals are jammed, and

$$\Pr\{\text{error}|J_{i-1},\bar{J}_i\} = \Pr\{\text{error}|\bar{J}_{i-1},J_i\} = \frac{1}{2}[1-Q(b,a)+Q(a,b)] \tag{13.69}$$

when either but not both are jammed, where $Q(\cdot,\cdot)$ is the Marcum Q-function[9] and

$$a = \sqrt{\frac{S}{2}}\left[\frac{1}{\sqrt{N_0}} - \frac{1}{\sqrt{N_0+N_e}}\right] \tag{13.70}$$

$$b = \sqrt{\frac{S}{2}}\left[\frac{1}{\sqrt{N_0}} + \frac{1}{\sqrt{N_0+N_e}}\right] \tag{13.71}$$

Equations (13.67) and (13.68) can be written in terms of the SNR given by

$$\gamma = \frac{S}{N_0} \tag{13.72}$$

and the JSR given by

[9] The Marcum Q-function is given by

$$Q(a,b) = \int_b^{\infty} xe^{\frac{x^2+a^2}{2}} I_0(ax)dx$$

where $I_0(\cdot)$ is the 0th order Bessel function of the first kind.

$$\xi = \frac{J_0}{S} \tag{13.73}$$

as

$$\Pr\{\text{error} \mid \bar{J}_{i-1}, \bar{J}_i\} = \frac{1}{2}\exp\left[-\gamma\right] \tag{13.74}$$

and

$$\Pr\{\text{error} \mid J_{i-1}, J_i\} = \frac{1}{2}\exp\left[-\frac{\gamma/\xi}{1/\xi + k\gamma}\right] \tag{13.75}$$

where

$$k = \frac{S_\text{n} R_i}{K R_\text{c}} \tag{13.76}$$

Likewise, (13.70) and (13.71) can be written

$$a = \sqrt{\frac{\gamma}{2}} - \sqrt{\frac{\gamma/\xi}{2\left(1/\xi + k\gamma\right)}} \tag{13.77}$$

$$b = \sqrt{\frac{\gamma}{2}} + \sqrt{\frac{\gamma/\xi}{2\left(1/\xi + k\gamma\right)}} \tag{13.78}$$

To include the effects of Rayleigh fading and lognormal shadowing, γ and ξ must be treated as r.v.s rather than constants, and the joint pdf $p(\gamma, \xi)$ must be determined and the conditional probabilities (13.69), (13.74), and (13.75) must be averaged over the range of the r.v.s. We can refine the conditional probabilities by making the following definitions [24]

$$\bar{\gamma} = \frac{4 \times 10^{\bar{m}_\text{d}/10}}{2\pi N_0} \tag{13.79}$$

which is the mean SNR,

$$\bar{m}_{\mathrm{d}} = \frac{4 \times 10^{m_{\mathrm{d}}/10}}{2\pi N_0} \tag{13.80}$$

which is the normalized mean with respect to the noise power spectral density. The conditional probabilities, including the effects of Rayleigh fading and log-normal shadowing are then

$$\Pr\{\mathrm{error}|\bar{J}_{i-1}, \bar{J}_i\} = \frac{1}{2} - \frac{5}{\sqrt{2\pi\sigma\ln(10)}} \int_0^\infty \frac{1}{\bar{\gamma}+1} \exp\left[-\frac{1}{2\sigma^2}\left(10\log_{10}\frac{\bar{\gamma}}{\bar{m}_{\mathrm{d}}}\right)^2\right] d\bar{\gamma} \tag{13.81}$$

$$\begin{aligned}&\Pr\{\mathrm{error}|J_{i-1}, J_i\}\\ &\quad = \frac{1}{2} - \frac{5}{\sqrt{2\pi\sigma\ln(10)}} \int_0^\infty \frac{1}{\bar{\gamma}+k\nu+1} \exp\left[-\frac{1}{2\sigma^2}\left(10\log_{10}\frac{\bar{\gamma}}{\bar{m}_{\mathrm{d}}}\right)^2\right] d\gamma_0\end{aligned} \tag{13.82}$$

where

$$\nu = \frac{\bar{m}_{\mathrm{d}}}{\bar{m}_{\mathrm{d}\varepsilon}} \tag{13.83}$$

with

$$\bar{m}_{\mathrm{d}\varepsilon} = \frac{4 \times 10^{m_{\mathrm{d}}/10}}{2\pi J_0} \tag{13.84}$$

$$\begin{aligned}&\Pr\{\mathrm{error}|\bar{J}_{i-1}, J_i\} = \Pr\{\mathrm{error}|J_{i-1}, \bar{J}_i\}\\ &\quad = \frac{1}{2} - \frac{5}{\sqrt{2\pi\sigma\ln(10)}}\\ &\qquad \times \int_0^\infty \frac{1}{\sqrt{(\bar{\gamma}+k\nu+1)(\bar{\gamma}+1)}} \exp\left[-\frac{1}{2\sigma^2}\left(10\log_{10}\frac{\bar{\gamma}}{\bar{m}_{\mathrm{d}}}\right)^2\right] d\bar{\gamma}\end{aligned} \tag{13.85}$$

For the special case of Rayleigh fading only with no shadowing $\sigma = 0$ so

$$\Pr\{\text{error}\,|\,\bar{J}_{i-1},\bar{J}_i\} = \frac{1}{2(\bar{m}_\text{d}+1)} \tag{13.86}$$

$$\Pr\{\text{error}\,|\,J_{i-1},J_i\} = \frac{k\nu}{2(\bar{m}_\text{d}+k\nu+1)} \tag{13.87}$$

$$\begin{aligned}\Pr\{\text{error}\,|\,\bar{J}_{i-1},J_i\} &= \Pr\{\text{error}\,|\,J_{i-1},\bar{J}_i\}\\ &= \frac{1}{2}-\frac{\bar{m}_\text{d}}{2\sqrt{(\bar{m}_\text{d}+k\nu+1)(\bar{m}_\text{d}+1)}}\end{aligned} \tag{13.88}$$

When there are H hopping channels used by the target network and K of those are jammed, where $0 < K \le H$, then the probability that a channel is jammed is given by K/H and the probability that a hop is not jammed is given by $1 - K/H$. Whether a jammed channel is used by the target during one time interval is independent of whether a jammed channel is used during the following interval when the jammer hops as well, which we assume here. Therefore the a priori probabilities of a channel being jammed are

$$\Pr\{J_{i-1},J_i\} = \left(\frac{K}{H}\right)^2 \tag{13.89}$$

$$\Pr\{J_{i-1},\bar{J}_i\} = \Pr\{\bar{J}_{i-1},J_i\} = \frac{K(H-K)}{H^2} \tag{13.90}$$

$$\Pr\{\bar{J}_{i-1},\bar{J}_i\} = \left(\frac{H-K}{H}\right)^2 \tag{13.91}$$

yielding the overall BER of

$$\begin{aligned}P_\text{e} = {}&\Pr\{\text{error}\,|\,J_{i-1},J_i\}\Pr\{J_{i-1},J_i\} + \Pr\{\text{error}\,|\,J_{i-1},\bar{J}_i\}\Pr\{J_{i-1},\bar{J}_i\}\\ &+\Pr\{\text{error}\,|\,\bar{J}_{i-1},J_i\}\Pr\{\bar{J}_{i-1},J_i\} + \Pr\{\text{error}\,|\,\bar{J}_{i-1},\bar{J}_i\}\Pr\{\bar{J}_{i-1},\bar{J}_i\}\end{aligned} \tag{13.92}$$

using (13.81)–(13.85) when both Rayleigh fading and log-normal fading are included and (13.86)–(13.88) when only Rayleigh fading is included.

As an example of these results, consider the case when

$$R_\text{c}/R_i = 31 \qquad H = 2 \qquad K = 1 \tag{13.93}$$

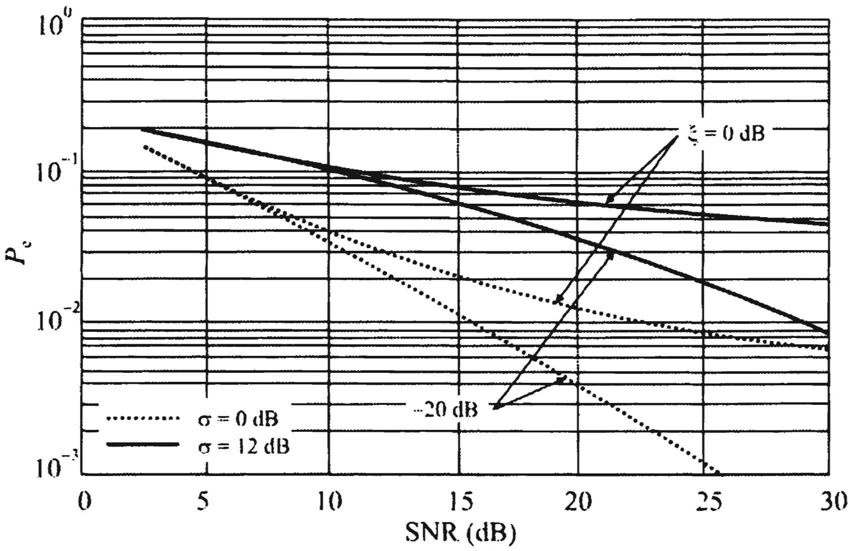


Figure 13.29 Hybrid receiver performance (R_c/R_i = 31, H = 2, K = 1). (*Source:* [23]. © IEEE 1991. Reprinted with permission.)

Figure 13.29 illustrates this case for Rayleigh fading when σ = 0 and with both Rayleigh fading and log-normal shadowing when σ = 12 dB, approximately the worst case of log-normal shadowing.

13.8 Concluding Remarks

We discussed jamming performance against narrowband and wideband targets when the communications channels involved exhibit fading characteristics in this chapter. Several scenarios were included where the fading can be Rayleigh, with no direct component; Ricean, where there is a discernable direct component; and Nakagami-*m*, which is a more general form of fading than the other two, and includes Rayleigh as a special case when $m = 1$.

In general it can be said that the target link fading has more effect on jamming performance than when the jammer link exhibits fading. This makes sense when we consider that the digital communication link must achieve a fairly low BER, say 10^{-3} or better, in order to be effective. On the other hand, the jammer only needs to create an environment where the BER is 10^{-1} or worse.

References

[1] Urkowitz, H., "Energy Detection of Unknown Deterministic Signals," *Proceedings of the IEEE*, Vol. 55, No. 4, April 1967, pp. 523–531.

[2] Kostylev, V. L., "Energy Detection of a Signal with Random Amplitude," *Proceedings IEEE International Conference on Communications*, New York, May 2002, pp. 1606–1610.

[3] Digham, F. F., M.-S., Slouni, and M. K. Simon, "On the Energy Detection of Unknown Signals over Fading Channels," *IEEE Transactions on Communications*, Vol. 55, No. 1, January 2007, pp. 21–24.

[4] Digham, F. F., M.-S., Slouni, and M. K. Simon, "On the Energy Detection of Unknown Signals over Fading Channels," *Proceedings IEEE International Conference on Communication*, Anchorage, May 2003, pp. 3575–3579.

[5] Oetting, J. D., "The Effects of Fading on Antijam Performance Requirements," *IEEE Journal on Selected Areas in Communications*, Vol. SAC-5, No. 2, February 1987, pp. 155–161.

[6] Poisel, R. A., *Modern Communications Jamming Principles and Techniques*, Norwood, MA: Artech House, 2004.

[7] Glenn, A. B., and G. Lieberman, "Performance of Digital Communications Systems in an Arbitrary Fading Rate and Jamming Environments," *IEEE Transactions on Communications Systems*, March 1963, pp. 57–68.

[8] Lieberman, G., "Spectra of Fading Signals," December 1961, pp. 254–260, 1958, (Unpublished report).

[9] Sonninschein, A., and P. M. Fishman, "Radiometric Detection of Spread-Spectrum Signals in Noise of Uncertain Power," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 28, No. 3, July 1992, pp. 654–660.

[10] Peterson, R. L., R. E. Ziemer, and D. E. Borth, *Introduction to Spread Spectrum Communications*, Upper Saddle River, NJ: Prentice Hall, 1995, pp. 328–329.

[11] Nakagami, M., "The *m*-Distribution—A General Formula of Intensity Distribution of Fading," in Hoffman, W. C., (Ed.), *Statistical Methods in Radio Wave Propagation*, Oxford: Pergamon 1960.

[12] Wojnar, A. H., "Unknown Bounds on Performance in Nakagami Channels," *IEEE Transactions on Communications*, Vol. Com-34, No 1, January 1986, pp. 22–24.

[13] Al-Hussaini, E. K., "Effects of Nakagami Fading on Antijam Performance Requirements," *Electronics Letters*, February 18, 1988, Vol. 24, No. 4, pp. 208–209.

[14] Abramowitz, M., and I. S. Stegun (Eds.), *Handbook of Mathematical Functions* (National Bureau of Standards, 1972), 6.2.2, p. 258.

[15] Abramowitz, M., and I. S. Stegun (Eds.), *Handbook of Mathematical Functions* (National Bureau of Standards, 1972), 6.2.1, p. 258.

[16] Simon, M. K., J. K. Omura, R. A. Scholtz, and B. K. Leavitt. *Spread Spectrum Communications Handbook*, Revised Ed., New York: McGraw-Hill, 1994.

[17] McGuffin, B. F., "Jammed FH-FSK Performance in Rayleigh and Nakagami-*m* Fading," *Proceedings IEEE MILCOM* 2003, pp. 1077–1082.

[18] McGuffin, B. F., "Distributed Jammer Performance in Rayleigh Fading," *Proceedings IEEE MILCOM* 2002, Anaheim, CA, Oct. 2002, pp. 669–614.

[19] Abramowitz, M., and I. S. Stegun (Eds.), *Handbook of Mathematical Functions* (National Bureau of Standards, 1972), Section 26.6.

[20] Abramowitz, M., and I. S. Stegun (Eds.), *Handbook of Mathematical Functions* (National Bureau of Standards, 1972), 13.2.5, p. 505.

[21] Abramowitz, M., and I. S. Stegun (Eds.), *Handbook of Mathematical Functions* (National Bureau of Standards, 1972), Section 5.

[22] Muammar, T., "Performance Evaluation of a Hybrid Spread Spectrum System in a Hostile Land Mobilé Radio Channel," *Proceedings 41st IEEE Vehicular Technology Conference, 1991: Gateway to the Future Technology in Motion*, May 19–22, 1991, pp. 108–113.

[23] French, R. C., "The Effects of Fading and Shadowing on Channel Reuse in Mobile Radio," *IEEE Transactions on Vehicular Technology*, Vol. 28, No. 3, August 1979.

[24] Yost, R. A., and R. H. Pettit, "Susceptibility of DS/FH Binary DBPSK to Partial and Full Band Barrage Jamming," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 17, No. 5, September 1981.

Chapter 14

Electronic Attack: UAS and Ground-Based

14.1 Introduction

Denying communications is the function of EA. This can be accomplished by interjecting more RF energy into the target receivers than the intended transmitter does. This chapter presents results of static calculations of jammer-effective range from a few configurations of jamming platforms. This range is the distance from the jammer within which the jammer will remain effective.

For effective communications the SNR at the receiver must be above some minimum level in the absence of a jammer. The SNR is

$$\gamma = \frac{P_{\mathrm{R}}}{N_{\mathrm{total}}} \tag{14.1}$$

which is frequently expressed in decibels. P_{R} is the power from the receive antenna. Without a jammer

$$N_{\mathrm{total}} = N_{\mathrm{noise}} \tag{14.2}$$

where

$$N_{\mathrm{noise}} = N_{\mathrm{external}} + N_{\mathrm{internal}} \tag{14.3}$$

as discussed in Chapter 5.

With a jammer present the total noise is expressed as

$$N_{\mathrm{total}} = N_{\mathrm{noise}} + J \tag{14.4}$$

where J represents the jammer power. Under many electronic attack scenarios $J >> N_{\text{noise}}$ and N_{noise} can be, and are, neglected in such calculations. In these cases the JSR = ξ is the quantity of interest, which is the reciprocal of (14.1).

As discussed in Chapter 7, modern digital forms of communication are easier to jam than the older, analog modulations. A JSR of 0–6 dB was necessary against analog radios. With digital modulations, creating a BER higher than about 10^{-2} is sometimes all that is necessary. This makes only 1 bit out of 100 in error. It has been shown that this can be accomplished at considerably less than 0 dB JSR. Here it is assumed that effective jamming occurs at zero JSR.

Probably the most significant advantage of EA from a UAS platform is the ability to overfly the target area and put the jamming energy where it is intended to be, rather than causing fratricide in friendly radios. This, of course, does not strictly apply to a truly nonlinear situation where friendly communications are in the same geographical area as the targets, but even then the jamming energy can be placed closer to the target than in standoff configurations. In the latter, the jammer has to emit energy over friendly communications before ever getting to the target area, always creating the possibility for fratricide. The results presented here are in the form of **jamming range**—that is, the range from the jammer within which, if the target receiver is located, the jammer will be effective at denying communications. Three jammer configurations were considered—the first is when the jammer is mounted in a UAS, the second is a ground-based vehicle configuration, and the last is a ground-based expendable configuration.

There are many jammer configurations possible. The ones included here are representative and are likely in land warfare scenarios. Similar analysis can be undertaken for any given configuration and for other types of propagation conditions—over water, for example, [1–10].

14.2 Signal Propagation at Long Ranges

For the purposes here the reflection model from Section 2.9 given by (2.43) will be used to calculate the signal and jammer powers at the receiver. Using this expression for the received power at all ranges will tend to underestimate the power at close ranges but beyond distance, d, given by

$$d = \frac{4h_T h_R}{\lambda} \tag{14.5}$$

where the following factors more accurately reflect the total power received: h_T = transmit antenna height AGL and h_R = receive antenna height AGL. As in

Chapter 15, since terrain is not considered, a circle defines the effective jamming region with the jammer at the center and with a radius defined by the calculated effective range.

14.3 Jamming ERP

When there is adequate room and the expense makes sense, antennas should be tuned according to the frequency at which they are used. Untuned antennas over 2 octaves (30–60 MHz, 60–120 MHz), the low-VHF range considered here, have substantial loss at either end of the frequency range. Losses of 15 dB or more would not be uncommon. Tuning these antennas matches the power amplifier output impedance to the antenna depending on the frequency. When tuned, losses of 3 dB across the frequency band are not unreasonable to expect. With only 3 dB loss due to the antenna, a 100W power amplifier produces an ERP of 50W, entirely reasonable for both the UAS and ground-vehicle configurations. With typical conversion efficiencies of 20% or so, that means that around 500W of prime power would be required for the system.

The expendable jammer case is different. Cost does not justify tuning the antenna. Furthermore, these jammers are powered by batteries that can deliver only so much power at a time and so much energy total. Thus, the expendable jammer must tolerate the antenna loss, and a power amplifier delivering 3W to an antenna with a 15 dB loss creates about 100 mW ERP. That is the power level used herein for the expendable jammer. It should be noted, however, that at midband the ERP would typically be much higher than this, so these results lean toward the worst case. The 3W power amplifier, at 20% conversion efficiency, would require about 15W of prime power. Two zinc carbon D-cells (normal flashlight batteries) can deliver 1.5V at about 4.5 amp-hours each at room temperature for a total of 13W for about 1 hour, sufficient for this jammer to operate for almost that hour.

The two mounted systems could be configured so that the frequency coverage could be from the low-VHF to beyond the PCS range, albeit different antennas may be required over the entire range. For the expendable jammer, one antenna probably could not provide the performance required to cover the whole range, so each jammer would probably be optimized (as much as possible) for only a portion of this range.

The expression for P_R (2.43) has a dependency on the height AGL of the jammer antenna, denoted here by h_J. It is the same as h_T in (2.43). For the UAS, h_J = 1,000m. For the mounted ground jammer h_J = 3m, while for the ground expendable jammer h_J = 1m.

14.4 Targets

Three types of targets are considered here. One of these is in the low-VHF range, the frequency range used by almost all military and paramilitary organizations worldwide. This target was at a frequency of 60 MHz, the middle of this range. The other target considered is the emerging PCSs expected to be used by these same forces in the future. In this case the handset is assumed to have an ERP of 0.5W, while the base station has an ERP of 10W.

The ability to jam a communication net is a function of RF link distances—from transmitter-to-receiver relative to the jammer-to-receiver link distance. For specificity, we will assume that the transmitter-to-receiver link distances are 5 km and 10 km, and that the antennas are omnidirectional in the horizontal plane. The ability to jam also depends on the relative power levels at the receiver. We will assume that in the VHF range, the PTT target transmit power is 10W ERP. The jammer power is 50W ERP. In the higher-frequency range of the PCS, P_J = 50W as well. When the target is the 10W PTT network, $h_T = h_R =$ 2m, reflecting ground-to-ground communications. When the target is the PCS handset, h_T = 10m, h_R = 2m and P_T = 10W. When the target is the PCS base station, h_T = 2m, h_R = 10m, and P_T = 0.5W.

14.5 RLOS

Radio waves traveling close to the Earth propagate for only relatively short distances as discussed in Chapter 4. These distances are referred to as the RLOS and due to refractive properties of the troposphere, do not typically correspond to the visual line of sight. Such refraction typically increases the radius of the Earth by a factor of one-third or so. For the heights used here, the RLOS is given in Table 14.1. Included are the RLOSs for the communicating nodes as well, which indicate that considering link distances of 5 and 10 km is reasonable.

Of course, there are propagation modes that allow waves to travel less than or longer than the RLOS—such modes are not considered here. The results presented herein then, can be considered nominal, with deviations either way possible.

14.6 UAS Jammer

Consider first the jamming range performance of a jammer mounted in a UAS. For the low VHF the jamming effectiveness regions are shown in Figure 14.1. Here the target transmitter power is 10W ERP. Note that the scale extends to ±100 km

Table 14.1 RLOS for the Antenna Heights Considered

Target	h_T (m)	h_R (m)	h_J (m)	RLOS (J→R) (km)	RLOS (T→R) (km)
PTT	2	2	Exp 1	5.5	6.4
	2	2	Gnd 3	7.1	6.4
	2	2	UAS 1,000	75	6.4
PCS Base	2	10	Exp 1	9.4	10.3
	2	10	Gnd 3	11.1	10.3
	2	10	UAS 1,000	78	10.3
PCS Handset	10	2	Exp 1	5.5	10.3
	10	2	Gnd 3	7.1	10.3
	10	2	UAS 1,000	75	10.3

from the jammer. In Figure 14.1(a) the link distance is 5 km, while in Figure 14.1(b) it is 10 km. The horizontal plane represented by the lighter color in Figure 14.1 represents the case when $\xi = 0$ dB, or the jamming power just equals the signal power at the target receiver. As discussed above, $\xi = 0$ is conservative when considering digital communications. Therefore, any time the plane is above the signal level contour, the jammer is ineffective—that is, the intended communication is getting through. On the other hand, wherever the J/S contour is above the plane shown, the jamming is effective and communication is denied. In this case, the jammer is effective out to a radius of the RLOS (75 km) from the UAS for both link distances.

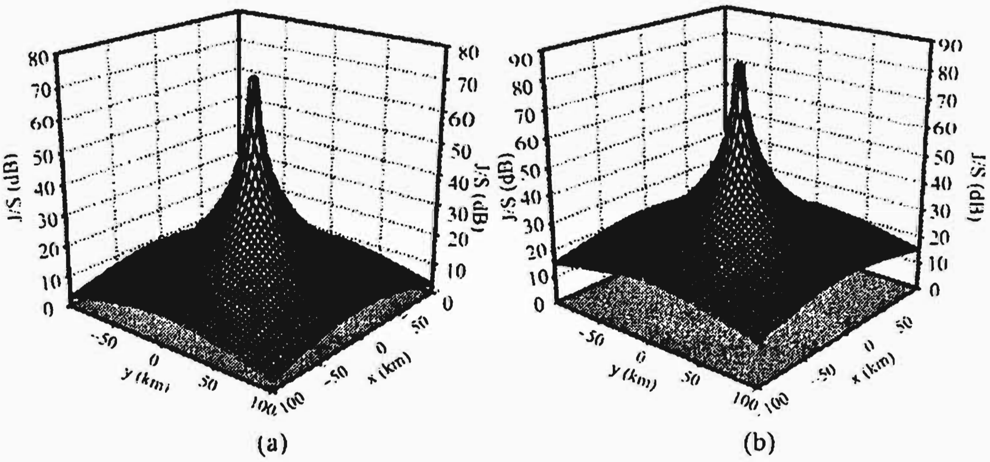


Figure 14.1 UAS jamming range performance when the target is a 10W PTT network and the link distance is (a) 5 km and (b) 10 km.

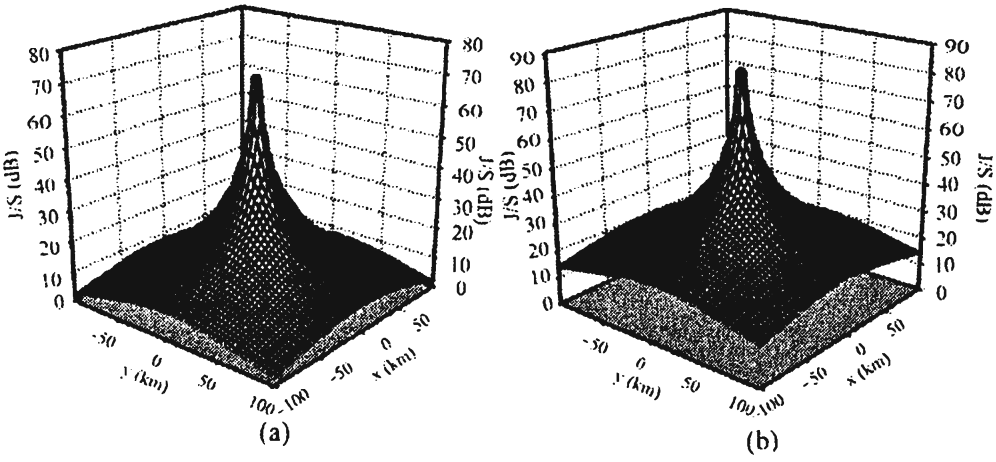


Figure 14.2 Jamming range performance for a UAS jammer when the target is a PCS base station and the link distance is (a) 5 km and (b) 10 km.

When the target is a PCS base station, the jamming range performance is as shown in Figure 14.2. Since jamming the base station implies P_T = 0.5W from the handset, the base station can be jammed out to RLOS for both link distances. When jamming the handset, P_T = 10W and the effective jamming range is reduced somewhat, as seen in Figure 14.3; however, the range is still RLOS for both link distances.

14.7 Ground-Based Jammer

The other mounted jammer system configuration considered is one that is ground-based. Again, a 50W ERP jamming waveform is assumed with an antenna that is omnidirectional in the horizontal plane.

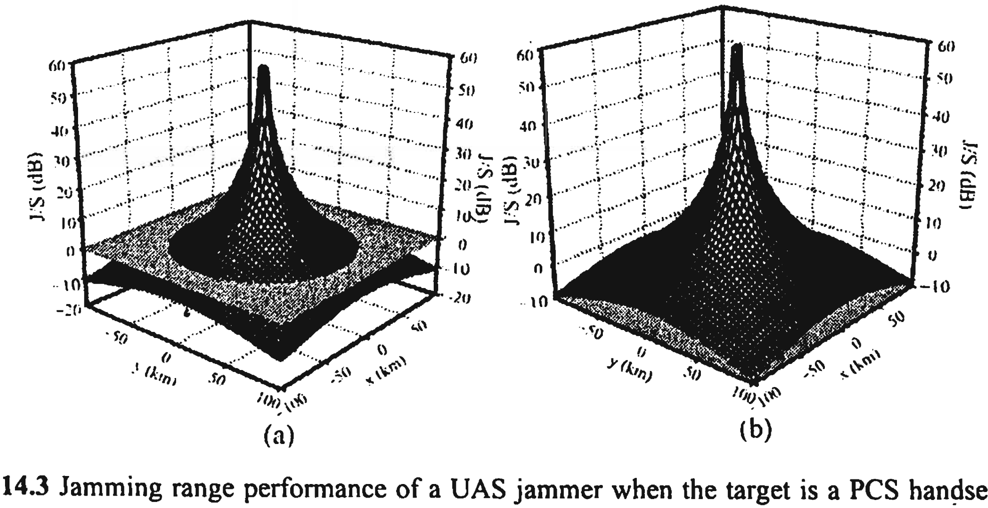


Figure 14.3 Jamming range performance of a UAS jammer when the target is a PCS handset and the link distance is (a) 5 km and (b) 10 km.

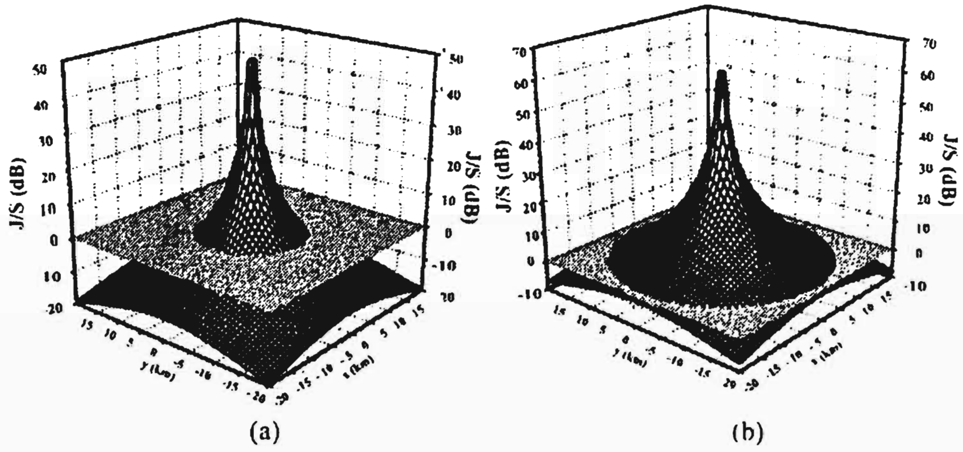


Figure 14.4 Ground jamming range performance when the target is a 10 W PTT network and the link distance is (a) 5 km and (b) 10 km.

When the link distance is 5 km for low-VHF targets, the effective jamming region is as shown in Figure 14.4(a). Note the scale change to ±20 km. The jammer is effective to a range of about 8 km in this case, which implies that the range is RLOS (7.1 km). At a link distance of 10 km that range is extended to about 17 km or so, again limited to RLOS at 7.14 km, as seen in Figure 14.4(b).

When the target is the PCS base station, the transmitter power is 0.5W and the effective jamming region is a circle with a radius of about 16 km as shown in Figure 14.5(a). For a 10 km link distance, the radius of the region is about 35 km. Thus, in both cases RLOS (11.1 km) limits the range.

The mounted ground-based jammer range when the target is the PCS handset is shown in Figure 14.6. Note the scale in this figure is ±20 km. Since the base station transmits more power than the handsets, the effective jamming range is less than for jamming the base station. With a 5 km link distance, the jamming range is

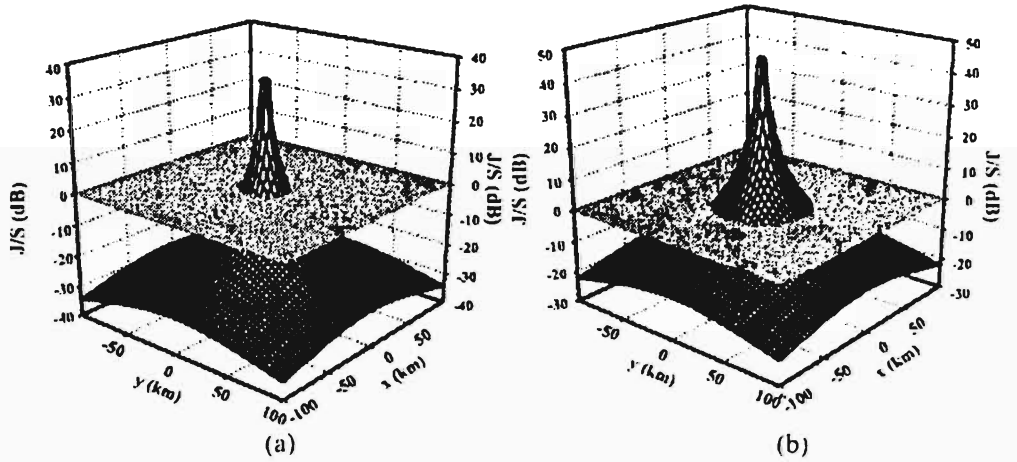


Figure 14.5 Ground jamming range performance when the target is a PCS base station and the link distance is (a) 5 km and (b) 10 km.

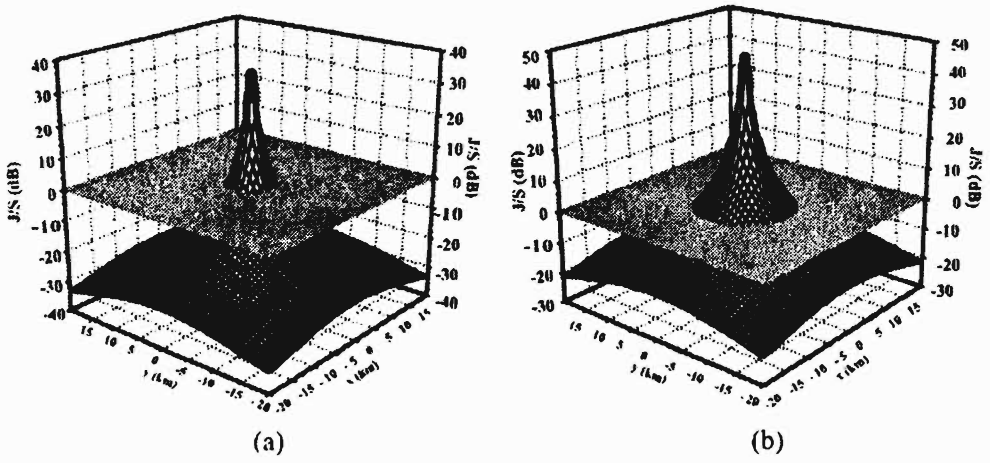


Figure 14.6 Ground jamming range performance when the target is a PCS handset and the link distance is (a) 5 km and (b) 10 km.

about 3 km or so as seen in Figure 14.6(a). With a 10 km link distance, it was about double this, to 6 km. Thus, in this case the jammer range is not limited by the RLOS, but by the jammer parameters.

It is highly desirable to be able to jam communication targets while the EA system is OTM. This implies significant restrictions on the type of antennas that can be used for this function. At 30 MHz, the wavelength is about 10m or so and a half-wave dipole or monopole would be about 5m long. While such an antenna could securely be mounted on a tactical vehicle, keeping it vertical while OTM would be a significant trick. Furthermore, a 5m antenna, sticking up vertically from a vehicle, violates many international traffic agreements, and would not clear highway overpasses in many countries, including the United States.

The jammer antenna height used here is h_J = 3m with a 3 dB loss. The 3 dB loss implies a certain antenna length and/or antenna tuning. Limiting h_J to 3m would allow for OTM operation, however.

14.8 Concluding Remarks

The results of the analysis of effective jamming ranges for the various system configurations are summarized in Table 14.2. In most cases the jamming range is limited only by the RLOS between the jammer and the target receiver.

The UAS jammer is quite effective, with a range of 75 km in all cases. This is the RLOS at the altitude at which the jammer was flown (1,000m AGL). This range would be longer if the UAS were to fly higher. A range of 75 km, however, is probably too long for most scenarios since unwanted fratricide would likely ensue at even 75 km.

Table 14.2 Jamming Ranges for the System Configurations Considered

	Target					
	10W PTT		**PCS base station**		**PCS handset**	
Link distance (km)	**5**	**10**	**5**	**10**	**5**	**10**
UAS	RLOS (75)	RLOS (75)	RLOS (78)	RLOS (78)	RLOS (75)	RLOS (75)
Ground mounted	RLOS (7.1)	RLOS (7.1)	RLOS (11.1)	RLOS (11.1)	3	6

When the 50W ERP jammer is in the ground-mounted configuration, again the limitation was the RLOS in most cases. It is assumed here that the jammer is configured for OTM operation so the jammer antenna height was compatible with a tactical vehicle that was moving—that is, the antenna was not elevated. A jammer antenna height of 3m is assumed. This height significantly restricts the RLOS range of the jammer.

As expected, the ground expendable jammer has the shortest effective range. With an ERP of only 100 mW, the range must be limited. In some cases it is less than 1 km and, in the case of the PCS handset with a 5 km link distance, cannot jam the target at all. (Of course, if the jammer were immediately adjacent to the handset, the latter would not be able to communicate.)

These results assume a sweeping narrowband jammer simulating a barrage jammer. Timing issues were not considered. The analysis in Chapter 7 showed that the timing could be arranged so that the jammer can be effective.

Jamming the handset end of the PCS link is the more difficult problem because the base station's transmit power is substantially higher than that of the handset and the height of the base station's antenna is substantially higher than that of a ground-based jammer. Although not considered here, directional antennas are normally used with such base stations. This would extend the range, reduce the amount of cochannel interference in the PCS system, and further limit the effectiveness of the jammer by null steering, for example, or employing the smart antenna concepts in Section 8.6.2.2.

References

[1] TRADOC Pamphlet 525-5, "The Objective Force, Operational Concepts, Organizational Design Constructs, and Materiel Needs Implications," U.S. Army Training and Doctrine Command, Fort Monroe, VA.

[2] U.S. Army Field Manual FM 100-6, *Information Operations*, Chapter 3 Operations, August 27, 1996.

[3] U.S. Army Field Manual FM 34-40-7, *Communication Jamming Handbook*, October 9, 1987.

[4] Neri, F., *Introduction to Electronic Defense Systems*, Norwood, MA: Artech House, 1991, pp. 337–416.

[5] Waltz, E., *Information Warfare Principles and Operations*, Norwood, MA: Artech House, 1998, Ch. 8.

[6] Adamy, D., *EW 101: A First Course in Electronic Warfare*, Norwood, MA: Artech House, 2001, pp. 177–222.

[7] Schleher, D. C., *Electronic Warfare in the Information Age*, Norwood, MA: Artech House, 1999, pp. 31–57.

[8] Mosinski, J. D., "Electronic Countermeasures," *Proceedings IEEE MILCOM*, 1992, pp. 191–195.

[9] "A Jamming Primer," *EW Design Engineers' Handbook*, Association of Old Crows, 1985; also contained in Hovanessian, S. A., *Radar System Design and Analysis*, Dedham, MA: Artech House, 1984.

[10] "Airborne ECM Tactics," *EW Design Engineers' Handbook*, Association of Old Crows, 1985; also contained in Van Brunt, L. B., *Applied ECM*, Vol. 1, 1978.

Chapter 15

Thin Jammers

15.1 Introduction

A thin jammer system consists of a set of small, relatively low-power jammers spread around a region. They are typically deployed on the ground with minimal antenna height and are battery operated so their duration and radiated power is limited. In most cases they are expendable so they must be inexpensive and therefore fairly unsophisticated. We present a discussion of thin jammer performance in this chapter.

A thin jammer system is characterized by relatively dumb jammers distributed throughout a region, with control exercised from a central location. The jammers are tasked according to some strategy, such as the closest jammer to the target is responsible for jamming the network. With modern networking technology, and ideas such as blue-force tracking, such tasking can be entirely reasonable and highly dynamic.

Since thin jammers such as these are typically placed on the ground, or at best, in trees close to the ground, we can expect that there will be significant propagation limitations. Therefore we include the effects of propagation constraints as discussed in Chapter 2. In addition, because of the placement limitations of the jammers, we can expect that fading effects will be significant. Being close to the ground with relatively short antennas, Rayleigh fading (no direct signal component) is the most likely type of fading encountered. Therefore we include these effects as discussed in Chapter 13.

Another configuration of thin jammers places such jammers on-board vehicles, where the primary purposes of the vehicles are not jamming platforms. The thin jammers are essentially parasites on these platforms and there are no EA operators uniquely associated with the jammer. C2 of the jammers is exercised from a remote location and targets are assigned dynamically. Thin jammers in such a configuration can implement higher radiated power than those that are battery powered. They can also be more sophisticated since, generally, they are not expendable. Another significant advantage is that they can be deployed closer to the target networks than equivalent, but standoff, jammer configurations.

For the analysis in this chapter, we will assume that the target networks consist of LPI radios. We will consider both fast- and slow-hopping transceivers as well as DSSS networks.

Since the jammers are assumed to be distributed among the target receivers, similar propagation characteristics would be expected on the transmitter to receiver as well as the jammer to receiver links. Thus similar fading characteristics are assumed for both links [1]. Furthermore, in highly irregular terrain, such as urban or mountainous, both links would experience time-varying multipath and, therefore, time-varying fading. In addition, realistic propagation conditions for antennas close to the Earth are considered [2, 3].

The simplest jamming waveform to use in such low-capacity jamming platforms is barrage jamming, where a portion (or all) of the frequency range of interest (such as the low VHF, 30–90 MHz) is jammed all the time. If a portion of the range is jammed, the jamming waveform may be moved in frequency with its portion of the spectrum, to eventually cover all of the range of interest.

Follower jamming is also possible, where a receiver (usually wideband compared to the total bandwidth of interest, but may be narrowband in some cases) senses the spectrum and determines where new energy is occurring. If the receiver is adequately desensitized, any new energy detected may be assumed to be from the target of interest, and that determines the frequency to be jammed next. Of course, some form of look-through is necessary to make such sensing measurements.

Another form of "smart jamming" uses pulsed signals. This technique also implies some form of ES receiver to make measurements in the spectrum and to determine the correct time to jam. In that case, jamming just the synchronization sequence may be possible [4], for example. Using such pulse techniques has the advantage of (1) using less energy to achieve the same JSR but only at the appropriate moment or (2) increasing thc peak jammer power for the same energy use, thereby increasing the jamming margin.

The performance of such distributed jammers is determined by the coverage areas achieved by an individual jammer over which some parameter, such as minimum JSR, is achieved. This coverage area herein is assumed to be a circle of radius R, implying that the antennas in use are omnidirectional in coverage (in the horizontal plane any.way), and that the effects of wavefront-disturbing objects are ignored (other than including the effects of statistical fading). The radius of the coverage circle from each jammer is an important system parameter because it determines, to a large extent, the cost of such a jamming system. The total system cost is proportional to R^{2}.

The fading characteristics are assumed to be Rayleigh because the antenna height and power available from each jammer are limited. A Ricean fading model would imply that a significant non-specular component to the jamming waveform is present. Propagation characteristics of both the target signal as well as the jamming signal are characterized by the R^n model discussed in Chapter 2, thus they experience loss at a rate of $1/r^n$, where r is the jammer-to-target range or the transmitter-to-receiver range, and $n \geq 2$.

Jammer performance improvement upper bounds with intelligent jamming are estimated by assuming the jammers are able to extract target signal information and calculate the improvement in jamming probability and/or extension in coverage range for both frequency follower and pulsed jammers. Performance improvement is greatest when each jammer coverage region contains a limited number of distinct communications subnetworks.

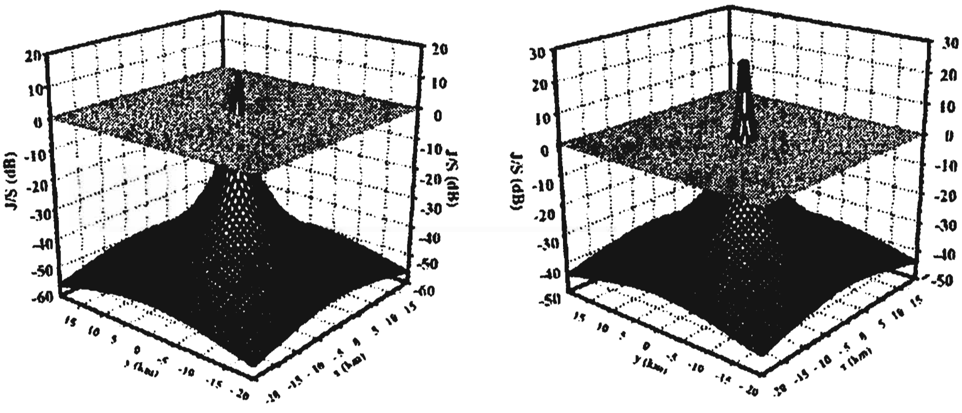
This chapter is structured according to the aforementioned categories. In the next section thin jammer performance is discussed without consideration of fading. That is followed by a presentation of the performance when fading on both links is included. After that, a discussion of frequency following jammer performance is offered. The chapter concludes with a section on the performance improvements made possible by pulsing the jammer, thereby increasing the radiated power.

15.2 Thin Jammers without Fading

We consider thin jammer performance in channels without taking into consideration the effects of fading in this section. Fading characteristics are included in the next section.

Thin jammers are of relatively low power and are expendable. Such jammers are inexpensive enough so that they need not be retrieved once deployed. They are normally battery operated so they perform EA for only a limited amount of time.

The antennas for these jammers are normally stowed until they are put into use. If hand-emplaced, the soldier emplacing the jammer could erect the antenna much like a retractable car antenna works—simply by pulling and extending the sections. If air-delivered, the antenna would have an automatic deployment mechanism. In either case, for the low-VHF range, the antenna would be about a half-wavelength long at midrange. Since the antenna would not be tunable, it would have considerable losses at the low and high ends. For the PCS frequency range, reasonably efficient antennas could be included. Because of the necessity of powering these devices with batteries and the inefficiency of the antenna as indicated in Section 14.3, such jammers would typically have an ERP of 100 mW or so averaged across the low-VHF frequency band.



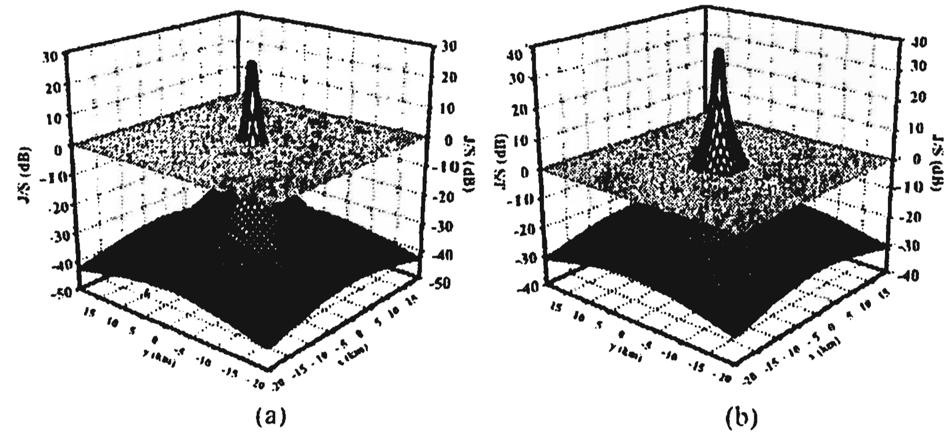
(a) (b)

Figure 15.1 Expendable jammer range when the target is a 10W PTT net and the link distance is (a) 5 km and (b) 10 km.

The results of simulating one of these jammers against the low-VHF targets are shown in Figure 15.1. With a 5 km target link the jamming range was less than 1 km while with a target link distance of 10 km, the range increases to about 2 km.

Against the PCS base station when $P_T = 0.5$ W, the jamming range was about 2 km when the link distance was 5 km as illustrated in Figure 15.2(a). When the link distance was extended to 10 km, about the typical edge of a PCS cell, the jamming range was about 5 km as seen in Figure 15.2(b).

When the target was the PCS handset, the jammer was ineffective at any range for the 5 km link as shown in Figure 15.3. For a 10 km link somewhat less than 1 km was the effective range. These results are summarized in Table 15.1.



(a) (b)

Figure 15.2 Expendable jammer range when the target is a PCS base station and the link distance is (a) 5 km and (b) 10 km.

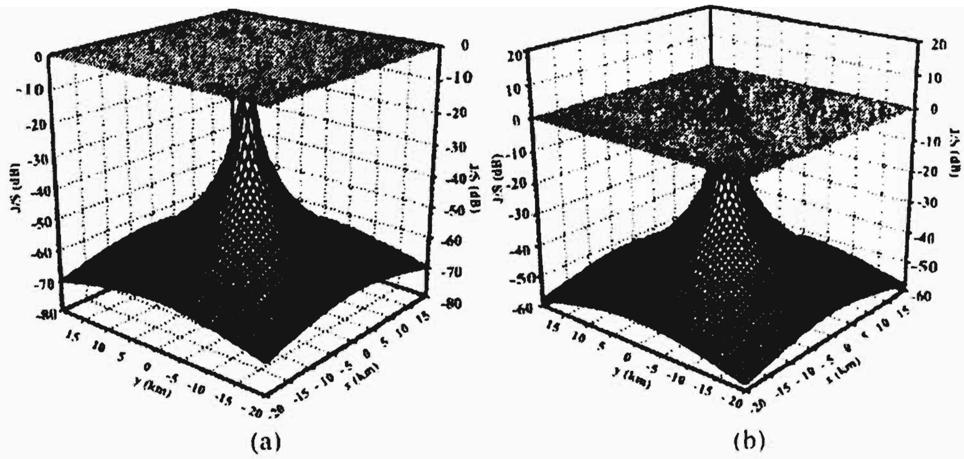


Figure 15.3 Expendable jammer range when the target is a PCS handset: (a) 5 km link distance and (b) 10 km link distance.

Table 15.1 Jamming Ranges for the System Configurations Considered

	Target					
	10W PTT		**PCS base station**		**PCS handset**	
Link distance (km)	**5**	**10**	**5**	**10**	**5**	**10**
Thin Jammer	<1	2	2	5	0	<1

15.3 Thin Jammers in Fading Channels

We include the effects on jamming performance when the links are degraded due to fading. Fading would most likely be encountered with thin jammers due to their close proximity to the ground.

15.3.1 Rayleigh Fading

As discussed in Chapter 13, the threat signal at the target receiver that is the subject of Rayleigh fading can be represented by

$$r(t) = \alpha_S(t)s(t) + \alpha_J(t)j(t) + n(t) \tag{15.1}$$

where $s(t)$ is the transmitted target signal, subjected to a delay, $\alpha_S(t)$ is the channel gain characterizing the fading experienced by the target signal, $j(t)$ is the jammer signal, $\alpha_J(t)$ is the channel gain experienced by the jamming signal, and $n(t) \sim \mathcal{N}[0, \sigma_\eta(t)]$. In general, $\alpha_S(t)$ and $\alpha_J(t)$ are complex so that

$$\alpha_S(t) = |\alpha_S(t)| e^{j\phi_S(t)} \tag{15.2}$$

and

$$\alpha_J(t) = |\alpha_J(t)| e^{j\phi_J(t)} \tag{15.3}$$

but since we are including only noise jamming waveforms in this analysis, the amplitude is sufficient and we can ignore the phases. Thus we will assume

$$\alpha_S(t) = |\alpha_S(t)| = \alpha_S \tag{15.4}$$

and

$$\alpha_J(t) = |\alpha_J(t)| = \alpha_J \tag{15.5}$$

which are unit-mean r.v.s with variances $\sigma^2_{\alpha_S}$ and $\sigma^2_{\alpha_J}$.

The power levels at the receiver due to the transmitter, jammer, and fading are characterized by their average values $\bar{S}$ and $\bar{J}$. The actual power levels received are given by $J = \alpha_J \bar{J}$ and $S = \alpha_S \bar{S}$ where α_J and α_S are given by (15.5) and (15.4), respectively. Furthermore, for both links,

$$F_A(\beta) = \Pr\{\alpha < \beta\} = 1 - e^{-\beta} \tag{15.6}$$

with pdf $p(\alpha) = e^{-\alpha}$.

15.3.2 Spread Spectrum Processing Gain

Whichever jamming method is used, for now it is assumed that the jammer transmits noise-like signals with power J in a bandwidth B_J yielding a jammer power density of $J_0 = J / B_J$ in W/Hz.

The target networks employ spread spectrum with processing gain specified by G_p approximated as

$$G_p = \frac{W_{ss}}{R_b} \tag{15.7}$$

where W_{ss} is the total spread spectrum bandwidth and R_b is the data rate. For FHSS with N_f channels each with bandwidth B_S Hz, $W_{ss} = B_J = N_f B_S$ so that $G_p = N_f B_S / R_b \approx N_f$. For DSSS, $G_p \approx R_c / R_b$, where R_c is the chip rate. In both cases

$$E_b / J_0 \approx G_p S / J \tag{15.8}$$

where S is the average signal power.

We assume that BBN jamming is the method of choice, both because of its simplicity in implementation and its performance against LPI waveforms. Therefore $B_J \geq B_S$, where B_S is the bandwidth of the target signal.

15.3.3 Jamming Measure of Effectiveness

Whether the jammer is successful at jamming the receiver is a function of the propagation conditions, the relative power levels of the target transmitter and the jammer, and the type of modulation employed. The jamming is successful if the SJR at the receiver satisfies

$$\frac{E_\text{b}}{J_0} < \gamma_\text{th} \tag{15.9}$$

or JSR = $\xi = J_0 / E_\text{b} > 1 / \gamma_\text{th}$ for the appropriate value of threshold, γ_th. The target network is assumed to be using some form of digital signaling with bit energy given by E_b. This threshold depends on the type of modulation employed by the target link, but is frequently between –10 dB to 0 dB [4–6].

We adopt the MOE for coded digital communication links of achieving a channel BER of 10^{-1} or higher [7, 8]. At this level, for many commonly used codes, punch-through occurs and the decoded BER rises above 10^{-1}. With this amount of decoded errors it is very difficult for the communication link to continue to be effective because on average, 1 bit out of every 10 is decoded in error.

Thus, from (15.9) and including (15.8) we have

$$\begin{aligned} \frac{J}{G_\text{p} S} &> \frac{1}{\gamma_\text{th}} \\ \frac{\alpha_\text{J} \bar{J}}{G_\text{p} \alpha_\text{S} \bar{S}} &> \frac{1}{\gamma_\text{th}} \\ \frac{\alpha_\text{J}}{\alpha_\text{S}} &> \frac{G_\text{p} \bar{S}}{\gamma_\text{th} \bar{J}} \end{aligned} \tag{15.10}$$

The probability of (15.10) occurring is given by

$$P_\text{J} = \Pr\left\{ \frac{\alpha_\text{J}}{\alpha_\text{S}} > \frac{G_\text{p} \bar{S}}{\gamma_\text{th} \bar{J}} \right\} = \frac{\gamma_\text{th} \bar{J} / G_\text{p} \bar{S}}{1 + \gamma_\text{th} \bar{J} / G_\text{p} \bar{S}} \tag{15.11}$$

A bit error occurs with probability given by the BER. The probability of a symbol error is given by the probability of jamming a symbol, and for the binary modulations considered here, is given by the BER. Therefore $P_\text{e} = P_\text{J}$, as given by (15.11).

15.3.4 Range-Loss[*]

As discussed in Chapter 4, the average jammer power at the target receiver is a function of the range between the jammer and the target receiver by the relation

$$\bar{J} = \frac{J_1}{r^n} \tag{15.12}$$

where J_1 is the nominal average power level 1m from the transmitter and n varies between 2 and 6, with 4 a typical value for communication signals close to the surface of the Earth [9]. Close to the antennas, $n = 2$ applies. However, beyond the turnover distance given by [10]

$$r_{\text{TO}} = \frac{4}{\lambda} h_{\text{T}} h_{\text{R}} \tag{15.13}$$

the range loss exponent changes to $n = 4$ (or higher). For example, if $h_{\text{T}} = 1\text{m}$, $h_{\text{R}} = 3\text{m}$ then at 100 MHZ, $r_{\text{TO}} = 4\text{m}$, almost assuredly requiring consideration of $n = 4$.

The targets are assumed to be located at uniformly distributed, random points inside a circle of radius R with the jammer at the center. The pdf for the location is therefore given by $p(R) = 1/\pi R^2$. The average probability of jamming is the expected value of (15.11) with the expectation over the circle. In polar coordinates this average probability is given by

$$\bar{P}_{\text{J}} = \frac{2}{R^2} \int_0^R \frac{Cr}{C + r^n} dr \tag{15.14}$$

where

$$C = \frac{\gamma_{\text{th}} J_1}{G_{\text{p}} \bar{S}} \tag{15.15}$$

The average target signal power, $\bar{S}$, is treated as a constant.

Equation (15.14) has known solutions for integer n. Some common values of use to us are

$n = 2$ [11–13]

$$\bar{P}_{\text{J}}(n = 2) = \frac{C}{R^2} \ln\left(1 + \frac{R^2}{C}\right) \tag{15.16}$$

$n = 3$ [13, 14]

$$\bar{P}_\text{J}(n=3) = \frac{2}{3}\frac{C^{2/3}}{R^2}\left\{\begin{aligned}&\frac{1}{2}\ln\left[\frac{R^3+C}{\left(R+C^{1/3}\right)^3}\right]\\ &\qquad+\sqrt{3}\tan^{-1}\left(\frac{2R-C^{1/3}}{\sqrt{3}C^{1/3}}\right)+\sqrt{3}\tan^{-1}\left(\frac{1}{\sqrt{3}}\right)\end{aligned}\right\} \tag{15.17}$$

$n = 4$ [14]

$$\bar{P}_\text{J}(n=4) = \frac{\sqrt{C}}{R^2}\tan^{-1}\left(\frac{R^2}{\sqrt{C}}\right) \tag{15.18}$$

Example [15]: Assume that $n = 4$ and the target link has average received power $\bar{S} = G_\text{RT}P_\text{Tx} / r^4$. The jammer provides the average jammer power at the receiver $\bar{J} = G_\text{RJ}P_\text{J} / r^4$. The variables G_RF and G_RJ are the gains and losses for the target communication link at the receiver in the direction of the target transmitter and the same for the jammer-receiver link at the receiver in the direction of the jammer, respectively. Similarly, P_Tx and P_J represent transmit powers into the antennas at the target transmitter and jammer, respectively. Let $G_\text{RJ} = G_\text{RT}$ (typically omni-directional in the horizontal plane), then both values can be set to 0 dB (= 1) and ignored. In addition, with this assumption, $J_1 = P_\text{J}$. In that case, if $P_\text{Tx} = 1$W, and the communication link range is 1 km, $\bar{S} = 1\times 1/(10^3)^4 = -120$ dBW. Assume that $R_\text{c} = 64R_\text{b}$ so that $G_\text{p} \sim 18$ dB, and the jamming threshold SNR is $\gamma_\text{th} = 0$ dB. Then $C = 1\times P_\text{J}/(64\times 10^{12}) = P_\text{J} + 102(\text{dB})$. Figure 15.4 shows the resulting average jamming probability as a function of jammer coverage radius, using jammer powers of 1, 10, and 100W.

The 100W jammer is comparable to the communications signal: the jammer has 20 dB higher power and the signal has 18 dB of processing gain, making the signals at the target receiver about the same. 100W is an unrealistically high power level for a small, battery-operated jammer, however. 100W might represent the jammer power available from a vehicle-mounted thin jammer. The architecture of such a configuration of thin jammers would be the same as for dismounted jammers, however, since there is more prime power available from a vehicle-mounted jammer, and, in the future, the vehicle could provide the vetronics, such configurations present some very attractive alternatives.

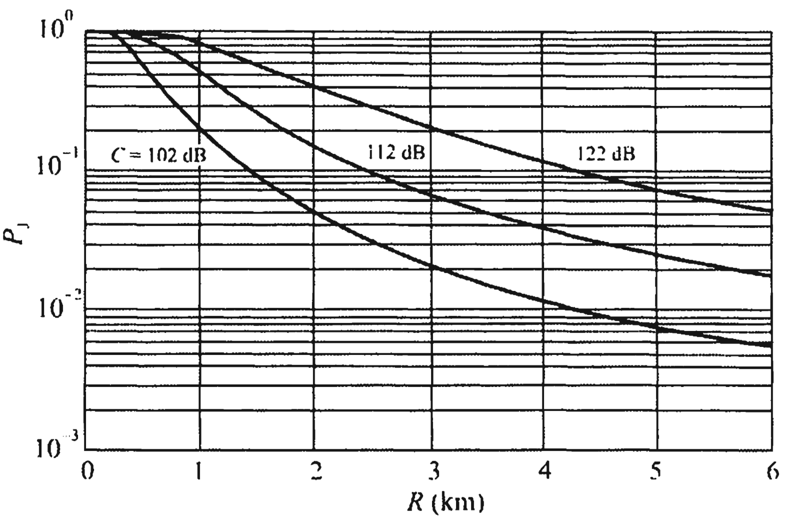


Figure 15.4 Average jamming probability as a function of coverage radius, parameterized by the normalized power ratio $C = \gamma_{\text{th}} J_1 / G_{\text{p}} \bar{S}$ with $1/r^4$ average range loss. The target link has 1W Tx power (equal to the weakest jammer), 1 km range, and 18.1 dB of processing gain. Combined range-independent gains and losses (antennas, height dependent propagation losses, etc.) are equal for the communication transmitter and jammer.

Using the $P_{\text{e}} \geq 10^{-1}$ criteria for digital communication links, we can see from Figure 15.4 that the 1W jammer achieves this point at about 1.3 km from the jammer. The 10 × jammer achieves the BER criteria at about 2.3 km, while the 100 × jammer gets out to about 4.3 km. Thus with the assumptions made here, the vehicle-mounted (100W) jammer has a range of just over 4 km.

Similar curves are shown in Figures 15.5 and 15.6 for the other propagation exponents as contained in (15.16)–(15.18). Note the change in the scale on the abscissa. The probability of jamming for the smaller exponents is substantially larger than when $n = 4$. The jamming signals propagate considerably further for small exponents.

The $n = 2$ curves correspond roughly to a UAS mounted thin jammer. Very long ranges are predicted. Of course, at substantially shorter ranges than those shown in Figure 15.5, the flat Earth assumption is violated. At typical UAS elevations, a total slant range of about 100 km would be about the limit. Nevertheless, comparison with Figure 15.4 indicates that the range-loss exponent has a remarkable effect on the coverage range.

15.4 Frequency Following Jammer

A frequency-hopping target signal usually only uses one channel at a time to transmit BFSK tones. Although it need not be the case, typically the two tones are

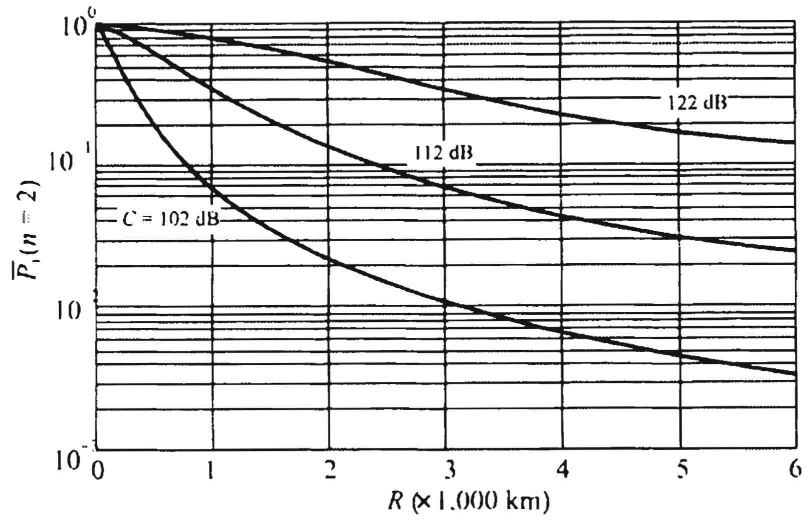


Figure 15.5 Probability of jamming success when the propagation exponent $n = 2$. All other parameters are the same as in Figure 15.4. (*Source:* [15]. © IEEE 2002. Reprinted with permission.)

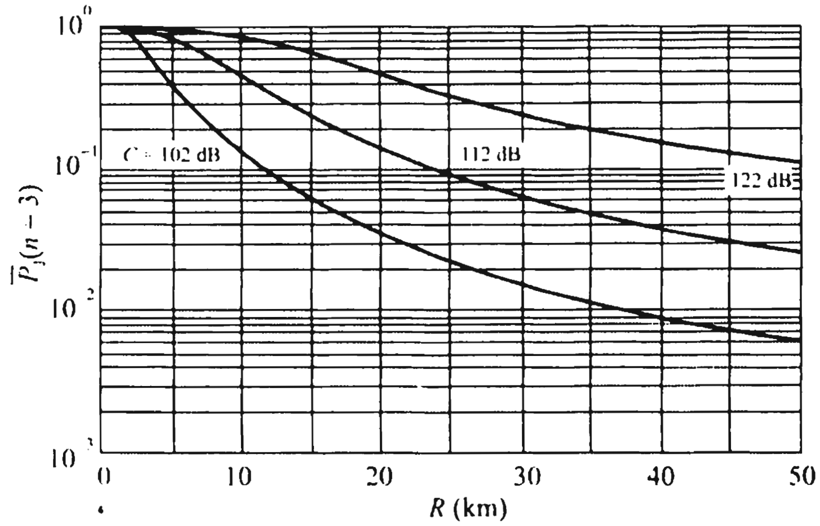


Figure 15.6 Probability of jamming success when the propagation exponent $n = 3$. All other parameters are the same as in Figure 15.4. (*Source:* [15]. © IEEE 2002. Reprinted with permission.)

sent within the same channel, say a channel 25 kHz wide in the VHF spectrum, and the two tones might be $f_0 \pm 5$ kHz where f_0 is the center of the channel. Therefore jamming is more effective if the jammer is able to concentrate its power in the channels actually in use. In this case G is replaced by $\mathcal{G}$, the ratio of bandwidth currently in use to the single-channel data rate

$$\mathcal{G} = \min\left(N_{\mathrm{f}}' \frac{B_{\mathrm{S}}}{R_{\mathrm{b}}}, G_{\mathrm{p}} \right) \tag{15.19}$$

where N_{f}' is the number of frequency channels in use inside the coverage region. For example, when both tones are sent in the same channel, $B_{\mathrm{S}}/R_{\mathrm{b}} \sim 1$. If $\mathcal{G}$ is known, it can be substituted directly for G_{p} in (15.15) for use in (15.16)–(15.18).

A true frequency following jammer selects one or more targets within its coverage area to jam, while ignoring others. In order for the jammer to follow a FHSS target from one frequency to the next, it must be able to somehow associate the next frequency to the last frequency used by the target network. This is a nontrivial task. Features that can be measured to help with that association include the angle of arrival of the target signal, although that is an indication of the direction of the transmitter not the target receiver that is to be jammed. Since thin jammers could be deployed within the same general region as the target networks, measuring the angle of arrival to the transmitter could point to a totally incorrect direction to the receiver. Furthermore, making such measurements in real time (the dwell time of a 100 hps frequency-hopping target is less than 10 ms) would typically require an antenna array at the jammer location.

Another possible feature is the dwell phase time. Many tactical frequency-hopping target radios use fixed dwell times, and the detection of a target at the correct time of day could be used. This implies a relatively sophisticated ES capability at the jammer, however, which runs contrary to the notion that such jammers should be unsophisticated and inexpensive.

For the considerations here, these limitations are not necessarily detrimental. Since the jammer coverage area is as small as it is, we can assume that the jammer does not distinguish one target from another within its coverage area and will place energy at all newly detected frequencies. Performance improvement would be expected in this case since energy is not wasted in portions of the spectrum in which there is no target. This, however, implies that signals outside of the circular region covered by a jammer are not detected. In this case, then, $B_{\mathrm{J}} = N_{\mathrm{f}} B_{\mathrm{S}}$ where N_{f} is the number of frequencies detected. These frequency channels need not be contiguous.

Of course an ES receiver is needed for spectrum monitoring to ascertain where the new target signal energy is. It must be able to determine frequencies. One implementation of such a receiver is a digital receiver that provides channelization with an FFT. Another possibility is a compressive receiver that scans, for example, 30–90 MHz in a matter of microseconds. A somewhat slower ES receiver, but fast enough for slow FFH targets in this same frequency range, is the swept superheterodyne with adequate instantaneous bandwidth discussed in Chapter 11. In addition to requiring an adequate ES receiver, jammer look-through is required. This is when the jammer is shut off momentarily so the receiver can measure the frequency spectrum.

Nevertheless, the jamming performance improvement that is possible by using frequency following can be estimated by assuming that the average number of target receivers is proportional to the coverage area, which we will again assume to be a circle of radius R with the jammer at its center. Let λ be the target receiver density, in receivers per unit area. Receivers in the same network use the same instantaneous frequency, reducing the number of channels in use. If an average network has N_N members, the average value of $\mathcal{G}$ is

$$\bar{\mathcal{G}} = \min\left(\frac{\lambda \pi R^2 B_S}{N_N R_b}, G_p\right) \tag{15.20}$$

Figure 15.7 shows how the jammer performance in Figure 15.4 will improve with frequency following. The dotted curve represents barrage jammer performance for the C = 112 dB (P_J = 10W, P_{Tx} = 1W) example. The solid curves show the

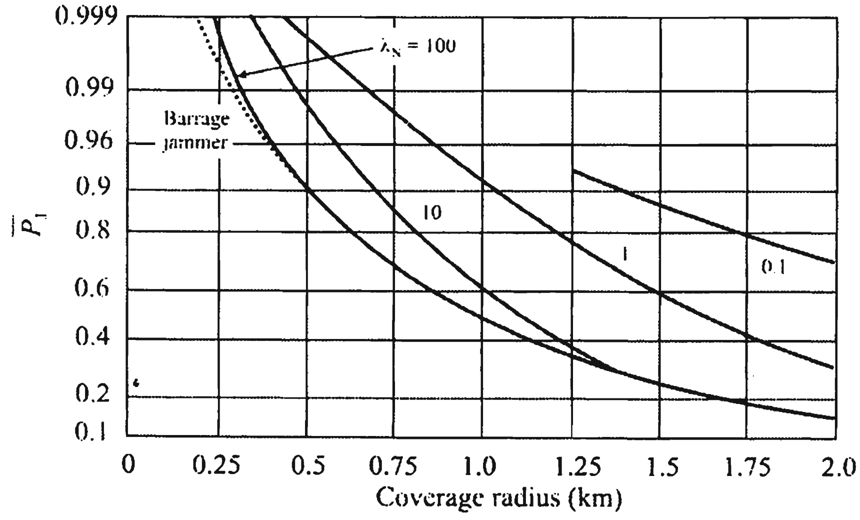


Figure 15.7 Frequency following jammer performance, C = 112 dB, W = 18.1, normalized signal bandwidth B_S/R_b, and range loss coefficient n = 4. Other parameters match Figure 15.4. (*Source:* [15]. © IEEE 2002. Reprinted with permission.).

frequency follower jammer parameterized by network density $\lambda_N = \lambda / N_N$. The average value of $\mathcal{G}$ was used, with processing gain G_p = 64 (18.1 dB) and $B_S / R_b = 1$. Notice that as the number of networks in the coverage area grows, jammer performance merges with that of the barrage jammer.

This performance improvement is an upper-bound on the actual improvement, because it assumes ideal frequency following, that is, all transmitted signals are accurately detected with no appreciable detection time, the jammer is able to transmit on only the frequencies in use, no matter how they are distributed, and, as mentioned, the jammer does not detect and attempt to jam signals transmitted to receivers outside its coverage area. In addition the jammer power at each receiver remains constant irrespective of the number of frequencies jammed.

15.5 Pulsed Jammer

As mentioned, digital target signals need not be jammed 100% of the time. It has been shown that creating a BER on the order of 10^{-1} can significantly degrade communications, causing approximately 1 bit in every byte to be in error. Therefore jammer efficiency can be improved by only turning the jammer on with a duty cycle of about 10%. For the type of jammers we are discussing here—distributed ones—the advantage of this is energy conservation which directly affects battery life. To the first order, the battery life can be extended 10 times.

We will use the average duty cycle for a given coverage radius. As before, the average number of receivers in the coverage region is $N_R = \lambda \pi R^2$. Each network is synchronized within itself and the networks are not synchronized with each other, so target transmissions on different networks may or may not overlap in time. If the target data for a single network has duty cycle δ, the average duty cycle to jam N networks is [15]

$$D(N) = 1 - (1 - \delta)^N \tag{15.21}$$

Using $N = \lceil \lambda \pi R^2 / N_N \rceil$ as the expected number of networks in a coverage region,[1] pulsed jamming can increase peak power by

$$G_D = \left[1 - (1 - \delta)^{(\lambda \pi R^2 / N_N)}\right]^{-\eta} \tag{15.22}$$

[1] $\lceil x \rceil$ denotes x rounded up to the nearest integer.

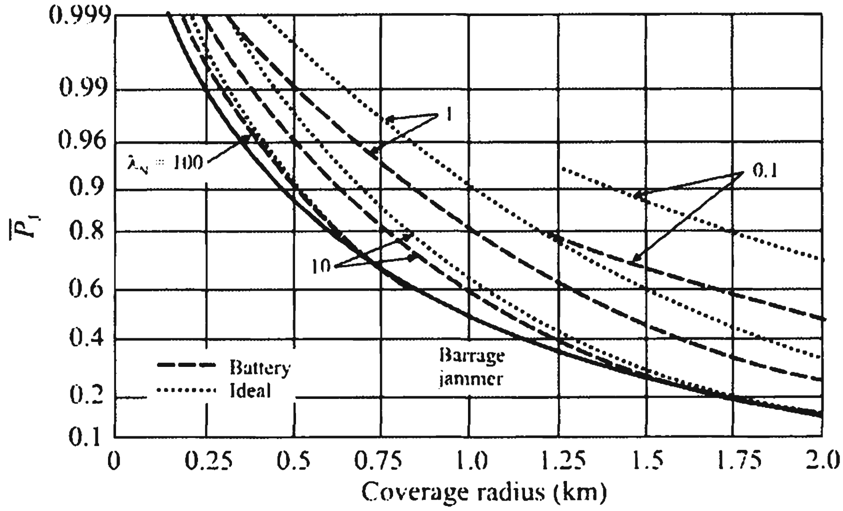


Figure 15.8 Pulsed jammer performance, C = 112 dB, single-network duty cycle δ = 1/64 = –18.1 dB, and range loss coefficient n = 4. Other parameters match Figure 15.4 (*Source*: [15]. © IEEE 2002. Reprinted with permission.)

where $\eta = 1$ for an ideal power source and $\eta \approx 2/3$ for batteries [15]. Figure 15.8 is a plot of jammer performance with different network densities λ_N. The single-network duty cycle was taken as $\delta = 1/64$ for comparison with the frequency following jammer. The pulsed jammer with an ideal power source (vehicle-mounted or fixed installation) performs virtually identically to the frequency follower. As expected, the pulsed jammer with a more realistic battery power source performs somewhat worse, especially when network density is low.

The result in this section is an upper-bound on achievable performance improvement because: it assumes the jammer will detect and synchronize to all networks in the coverage region; the jammer will not detect and synchronize to any networks operating outside the coverage region; and the power source is matched to each duty cycle, even though the duty cycle is variable.

15.6 Concluding Remarks

An evaluation of performance criteria for a thin jammer system was presented in this chapter. The first section considered such performance without taking into consideration the·fading that can occur in communication channels. The second portion included the effects when both the jammer and target signal are degraded by Rayleigh fading at the target receiver. The information provided can be used to determine how densely thin, distributed jammers must be deployed, and, hence, system costs. Propagation losses more severe than line-of-sight (n = 2) are more realistic in the near-ground environment because of the short antenna heights typical of distributed jammers [9].

We also considered the gross jammer performance improvements possible by implementing some forms of smart jamming. In particular, frequency following and time synchronized pulsed jamming were considered. Several simplifying assumptions were used to analyze these jammers, so the results provided should be considered only as upper bounds. Nevertheless, indications are that considerable improvements in jamming performance are possible. Frequency follower and pulsed jammer performance were essentially identical if the pulse duty cycle is the inverse of the spread spectrum processing gain and peak power scales linearly with duty cycle, i. e., average power is constrained. Using more realistic models of the peak power to duty cycle relationship for battery-operated jammers, frequency following was shown to provide superior performance.

Deployment of such thin jammers could be accomplished in several practical ways, such as hand-emplaced, artillery emplaced, or dropped from an aircraft such as a UAS, to name a few. Such jammer platforms have limited antenna height and limited prime power, the latter being furnished by an onboard battery with limited energy stores.

References

[1] Proakis, J. G., *Digital Communications*, New York: McGraw-Hill, 1983.

[2] Parsons, J. D., *The Mobile Radio Propagation Channel*, London: Pentech Press, 1992.

[3] Hess, G. C., *Handbook of Land Mobile Radio System Coverage*, Norwood, MA: Artech, 1998.

[4] Poisel, R. A., *Modern Communications Jamming Principles and Techniques*, Norwood, MA: Artech House, 2004.

[5] Torrieri, D. J., *Principles of Secure Communications Systems*, Norwood, MA: Artech House, 1985.

[6] Simon, M. K., J. K. Omura, R. A. Scholtz, and B. K. Levitt, *Spread Spectrum Communications*, Vol. I, Rockville, MD: Computer Science Press, 1985.

[7] Poisel, R. A., *Modern Communications Jamming Principles and Techniques*, Norwood, MA: Artech, 2004, Chapter 3.

[8] Nicholson, D. L., *Spread Spectrum Signal Design LPE and AJ Systems*, Rockville, MD: Computer Science Press, 1988, Chapter 5.

[9] Poisel, R. A., *Modern Communications Jamming Principles and Techniques*, Norwood, MA: Artech House, 2004, Chapter 2.

[10] Poisel, R. A., *Modern Communications Jamming Principles and Techniques*, Norwood, MA: Artech House, 2004, p. 26.

[11] Abramowitz, M., and I. A. Stegun, *Handbook of Mathematical Functions*, New York: Dover, 3.3.22, p. 121.

[12] Weast, R. C., (Ed.), *Handbook of Chemistry and Physics*, Cleveland: The Chemical Rubber Company, 1964, 65, p. A-129.

[13] http://www.sosmath.com/tables/integral/integ8/integ17.html, #2, accessed March 2008.

[14] http://www.sosmath.com/tables/integral/integ16/integ17.html, #2, accessed March 2008.

[15] McGuffin, B. F., "Distributed Jammer Performance in Rayleigh Fading," *Proceedings IEEE MILCOM* 2002, pp. 669–674.

Appendix A

Probability and Random Variables

A.1 Introduction

Communication signals represent probabilistic processes. Only in the simplest of cases can a communication signal be considered deterministic. A probabilistic process is one that can only be described with probability terms—the exact nature of such a process is not known a priori. A deterministic process, on the other hand, can be described exactly. Therefore it is useful for understanding communication signals and EW systems to have a basic understanding of random variables and probability.

Basic information on probability and random variables is included in this appendix. It certainly is not a detailed discussion, but it should be enough to understand what is necessary for the material herein.

A.2 Means, Expected Values, and Moment Functions of Random Variables

Given a set of N numbers $S = \{X_1, X_2, \ldots, X_N\}$, it is sometimes necessary to add all or some of these numbers together. A shorthand notation for representing this is given by Σ. Thus,

$$\sum_{i=1}^{N} X_i = X_1 + X_2 + \cdots + X_N \tag{A.1}$$

The *average*, or *mean*, value of this set of numbers is one number that is used in a sense to characterize the set that indicates where the "middle" is of the set of numbers. Herein it is denoted as μ. It is given by

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} \tag{A.2}$$

It is also known as the arithmetic mean of the set.

Suppose that X_1 is repeated n_1 times, X_2 is repeated n_2 times, ..., and X_N is repeated n_N times. Then the mean can be calculated as

$$\mu = \frac{\sum_{i=1}^{N} n_i X_i}{N} \tag{A.3}$$

where, of course, $n_1 + n_2 + \ldots + n_N = N$.

A characteristic of the set of numbers, which measures the "dispersion," or "spread" of the numbers around the mean value is the *standard deviation* of the set. It is denoted herein by σ and is given by

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}} \tag{A.4}$$

σ^2 is known as the *variance*. Again, if there are repeated values in the set, then the standard deviation can be computed as

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} n_i (X_i - \mu)^2}{N}} \tag{A.5}$$

Note that the standard deviation can be computed as

$$\sigma = \sqrt{\mu_S - \mu^2} \tag{A.6}$$

where μ_S is the mean of the set of numbers S_S consisting of the squares of the original set, namely, $S_S = \{X_1^2, X_2^2, \cdots, X_N^2\}$.

Generalizing the above, the *r*th *moment about zero*, μ_r, of the set of numbers is given by

$$\mu_r = \frac{\sum_{i=1}^{N} X_i^r}{N} \tag{A.7}$$

Similarly, the *r*th *moment about the mean* is given by

$$\mu_r = \frac{\sum_{i=1}^{N} (X_i - \mu)^r}{N} \tag{A.8}$$

The probability of a number, denoted p_i, for number X_i of the set S is given by the fraction of N corresponding to X_i (Kosko's the whole contained in the part [1]). Thus, in the above notation, $p_1 = n_1/N$, $p_2 = n_2/N$, and so forth. Note that by definition $\Sigma p_i = 1$.

The RMS is a measurement of the offset of the average of the set from zero as well as the spread around this average value. If μ is the average value, N is the number of data points, and the data values are denoted by X_i, then

$$\text{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2} \tag{A.9}$$

The *root sum squared* (RSS) is sometimes used to specify measurement errors. It is similar to the RMS calculation except that the mean value is not removed. Thus, it specifies more of the actual errors involved in measurements. The RSS is given by

$$\text{RSS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} X_i^2} \tag{A.10}$$

X is called a *random variable* (r.v.) if it can take on one of the values in a set of numbers (the set could be infinite in size) at one point (for example, at time t_n) and another value at another point, and the values it takes can only be described probabilistically. That means that ahead of time it is not known for certain what value X will take on at some future point. The expected value, or simply expectation, of a number X_i that has an associated probability of occurring in a given set is given by $\mathcal{E}\{X_i\} = p_i X_i$. If X is an r.v. that can take on the values from the set $S = \{X_1, X_2, \ldots, X_N\}$ where each X_i has associated probability of occurrence $p_1, p_2, \ldots, p_N$, then the expected value of X is given by

$$\mathcal{E}\{X\} = \sum_{i=1}^{N} p_i X_i \tag{A.11}$$

Note that with the previous notation with the set S, $p_i = n_i / N$. Substituting this into (A.11) yields

$$\mathcal{E}\{X\} = \sum_{i=1}^{N} \frac{n_i}{N} X_i \tag{A.12}$$

which is the mean value of the numbers as indicated previously.

A.3 Probability

Suppose a die is rolled and the results of each roll is a number $x \in \{1,2,3,4,5,6\}$ where $\in$ means "is an element of " and "{ }" denotes a set. If the die is rolled 1,000 times and some x, say, $x = 5$, turns up 200 times, then it is said that the probability of the occurrence of $x = 5$ in this experiment is 200/1,000 = 0.2.

Suppose that a card is pulled from a fair deck of cards (fair = not marked and all cards present). Then since there are 52 cards in a deck, what is the probability of drawing the jack of clubs? It is $1/52 \approx 0.019$ since there is one such card in the deck. What is the probability of drawing an ace? It is $4/52 \approx 0.077$ since there are four aces in the deck.

A process that has a random output x is called a *stochastic process* and x is referred to as an r.v. Such variables are described by their statistics, because a priori, we do not know the exact value the variable will take in any given instance. There are many instantiations of r.v.s in communication theory. The amplitude of an AM signal corrupted by noise is an r.v. The demodulated bit in a digitally modulated signal that is corrupted by phase noise is an r.v.

A.3.1 Conditional Probability

Suppose in a card experiment the question is the probability of drawing aces from a fair deck of cards. Furthermore suppose that on the first draw an ace has been drawn. What is the probability that a second ace will be drawn on the second try? This situation is described by *conditional probabilities*. Bayes, an eighteenth-century English clergyman, was the first person known to examine conditional probabilities. The probability of an event e occurring given that event a has already occurred is denoted as $\Pr(e|a)$. This conditional probability is given by the expression

$$\Pr(e|a) = \frac{\Pr(a,e)}{\Pr(a)} \qquad \text{(A.13)}$$

The numerator on the right side in this expression, $P(a, e)$, is the probability of both events a and e occurring. Rearranging this expression

$$\Pr(a,e) = \Pr(e|a)\Pr(a) \qquad \text{(A.14)}$$

If e and a are independent events, neither depends on the occurrence of the other. Therefore the probability of occurrence of e does not depend on whether a occurs or not. Thus $\Pr(e| a) = \Pr(e)$ and

$$\Pr(a,e) = \Pr(e)\Pr(a) \qquad \text{(A.15)}$$

In general,

$$\Pr(e,a) = \Pr(e)\Pr(a) - \Pr(e \,\text{or}\, a) \qquad \text{(A.16)}$$

Since $\Pr(e, a) = \Pr(a, e)$ then

$$\Pr(e|a)\Pr(a) = \Pr(a|e)\Pr(e) \qquad \text{(A.17)}$$

so that

$$\Pr(e|a) = \frac{\Pr(a|e)\Pr(e)}{\Pr(a)} \qquad \text{(A.18)}$$

In words, this expression says that the probability of e occurring given that a has already occurred is given by the ratio of the product of the probability of a occurring, given that e has already occurred times the a priori probability of e occurring at all, to the a priori probability of a occurring at all.

Two events a and e are *mutually exclusive* if the occurrence of one means that the other did not occur. In that case $\Pr(a, e) = 0$ and $\Pr(e|a) = \Pr(a|e) = 0$.

A.3.2 Random Variable Distribution and Density Functions

Suppose that you have a pair of dice. Furthermore, you do not know if the dice are "fair" in the sense that any one of the 6 × 6 = 36 combinations of numbers could occur equally. Now suppose you throw those dice 1,000 times and record the sum of the dice that occurred on each throw. Then you plot these results on a graph, with the results as shown in Figure A.1. The lines on the left are the experimental results while those on the right are the theoretical results if the dice were perfectly fair. Could you then say with some degree of confidence that the dice are "fair"? Probably not (the actual results are clearly skewed to the low side indicating one or more of the die is weighted low), but that is not the point here.

Now suppose that instead of plotting the number of occurrences of each sum themselves, the numbers are "normalized" so that each "times number occurred" represents a fraction of the total times (1,000) the dice were thrown. That leads to a similar graph as shown in Figure A.2 (plots of the ideal results have been removed for clarity). A graph such as this is known as a *probability density function* (pdf) for the discrete random variable "number." "Number" is discrete because it can take on only specific values. Note that the "area" under these bars totals 1, because $n_1/N + n_2/N + \ldots + n_N/N = (n_1 + n_2 + \ldots + n_N)/N = N/N = 1$.

If the random variable is allowed to take on continuous values, then a similar result ensues, except that the histogram is now a continuous curve. The area under this

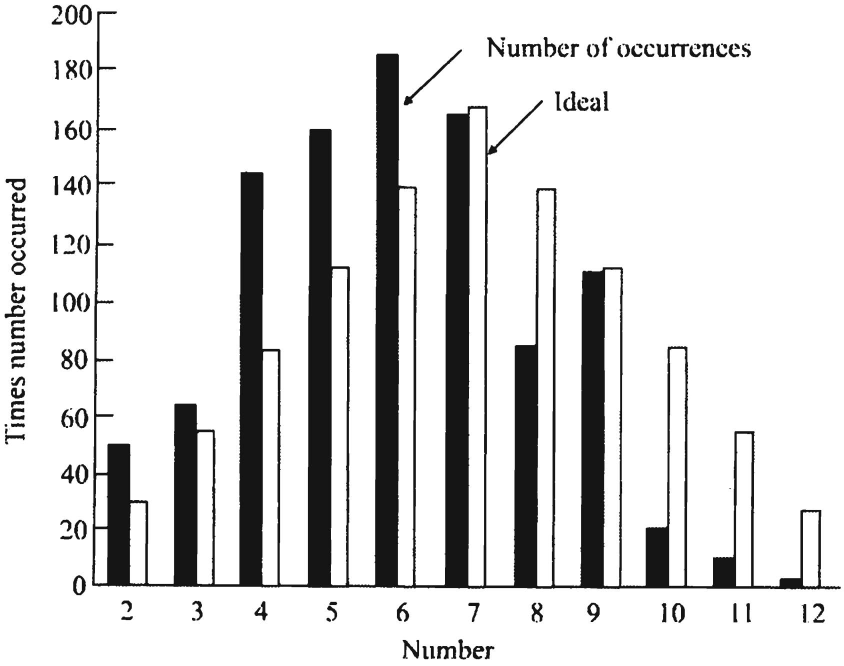


Figure A.1 Densities for the experiment.

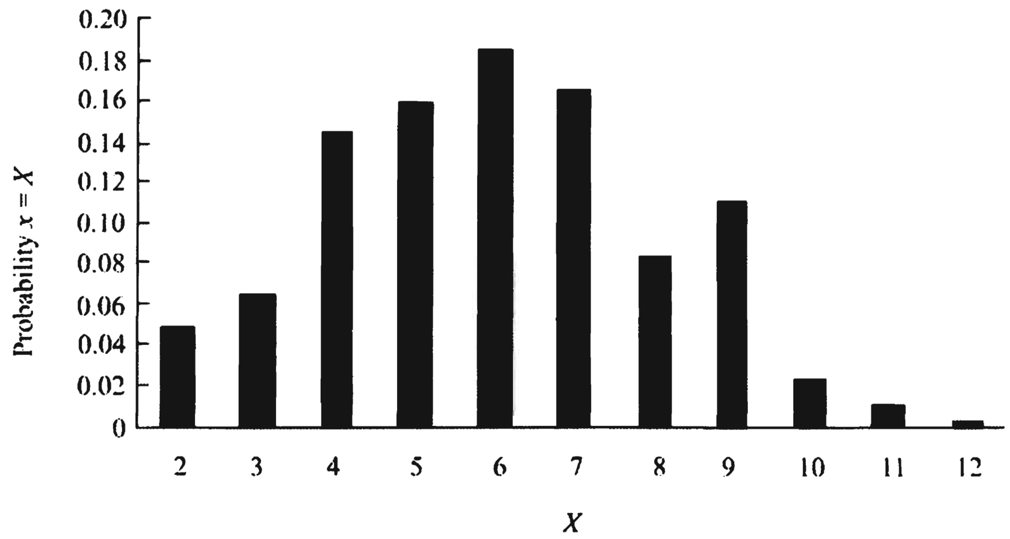


Figure A.2 Experiment results that have been normalized to the total number of trials.

$$f_X(x) = \frac{dF_X(x)}{dx} \qquad \text{(A.19)}$$

curve corresponds to probability mass and the area between two values of X, say, X_1 and X_2, corresponds to probability $(X_1 < X < X_2)$, as shown in Figure A.3.

For both discrete and continuous random variables there is a function known as the *cumulative distribution function* (cdf) which represents the probability that the r.v. x is less than some specified value, say, X. For the discrete case above, it is shown in Figure A.4. Since the continuous density function as shown above is based on the same data as the discrete density function, the distribution function

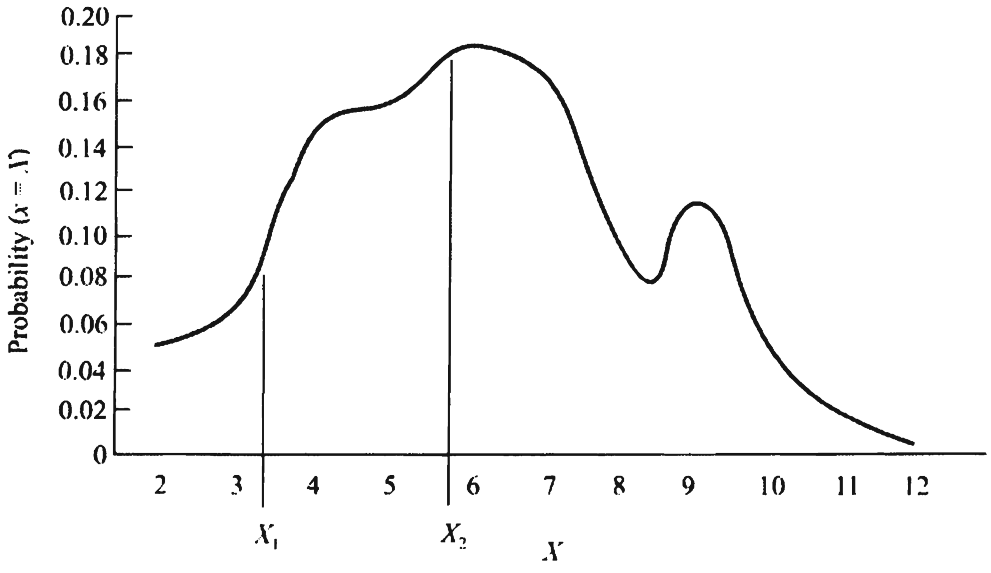


Figure A.3 Continuous density function.

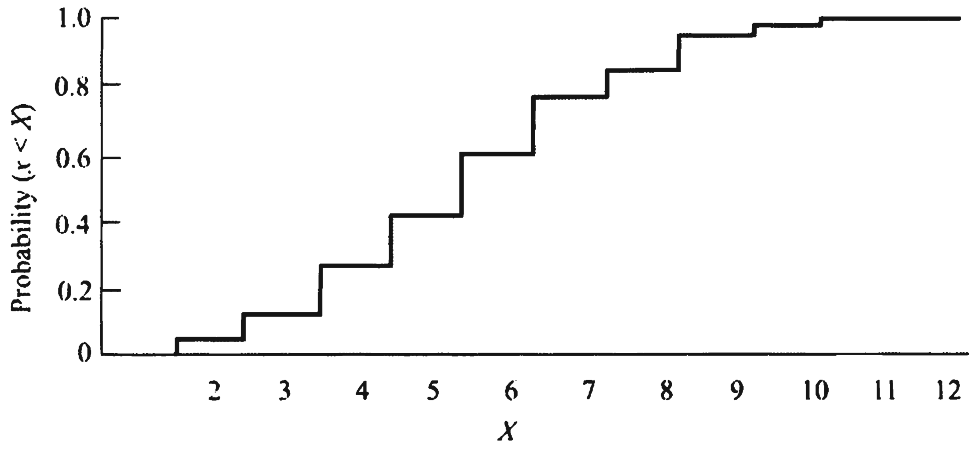


Figure A.4 Discrete probability distribution function.

for the continuous case here would look the same as Figure A.5. In mathematical terminology, the density function $f_X(x)$ and distribution function $F_X(x)$ are related by the equations

$$f_X(x) = \frac{dF_X(x)}{dx} \tag{A.20}$$

and

$$F_X(x) = \int_{-\infty}^{x} f_X(\xi)d\xi \tag{A.21}$$

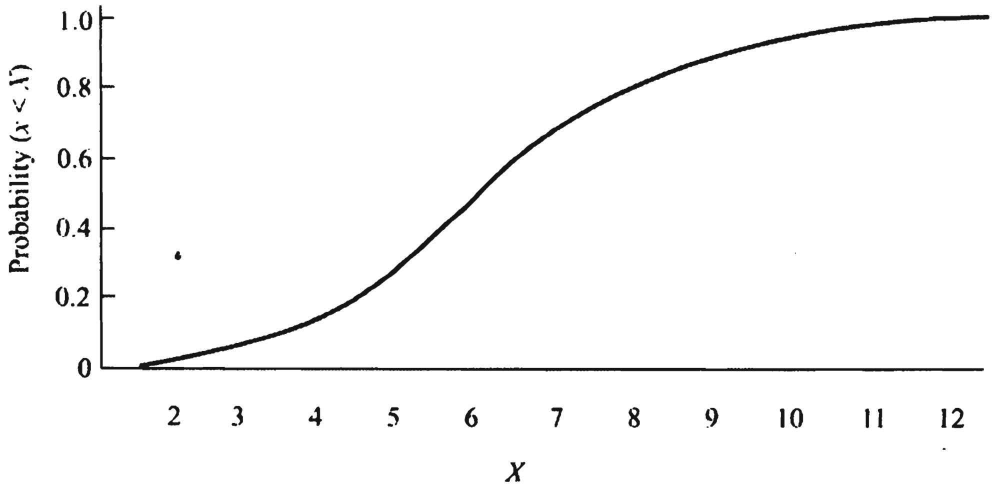


Figure A.5 Continuous probability distribution function.

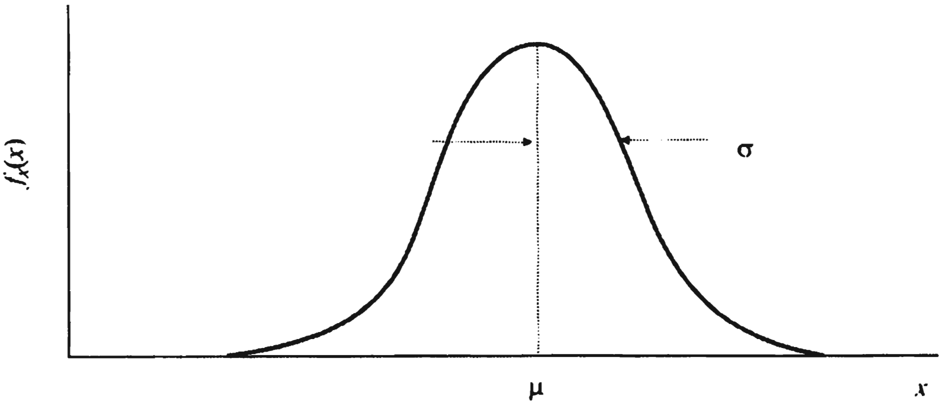


Figure A.6 Gaussian, or normal, probability density function.

That is, $F_X(x)$ is given by the area under the density function from $-\infty$ to x, and $f_X(x)$ is the slope of $F_X(x)$ at $x = X$.

A.4 Gaussian Density Function

One density function of particular interest to the analysis of communication signals and EW systems is the *Gaussian*, or *normal*, density function. Gauss was a 17th-century mathematician who discovered, among other things, that there is a probability density function that approximates the behavior of random variables in many circumstances. This is shown in Figure A.6. To interpret this function, like all probability density functions, the abscissa represents the values of a random variable, in this case denoted as x. Then the probability that x falls between x_1 and x_2 is given by

$$\Pr(x_1 < x < x_2) = \int_{x_1}^{x_2} f_X(x)dx \tag{A.22}$$

The average, or mean, amplitude is denoted as μ and the standard deviation is given by σ. This distribution is often used to describe the amplitude of noise signals. When it is, then the power in the noise signal is given by σ^2.

A random variable described by such a density function is frequently encountered in practice. For example, it describes reasonably well the noise voltage and current associated with resistors. The density is given by the equation

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \tag{A.23}$$

where μ is the mean and σ is the standard deviation.

The density function is shown in Figure A.6, while the distribution function is shown in Figure A.7. In both of these figures $\mu = 0$ and $\sigma = 1$, but the results are easily generalized to other mean and standard deviation values.

For the Gaussian density, 68.27% of the probability mass lies within $\pm 1\sigma$ of the mean value, while 95.45% of the probability mass lies within $\pm 2\sigma$ of the mean, and 99.73% of the probability mass lies within $\pm 3\sigma$ of the mean.

A.5 Kurtosis

The *kurtosis* is a measure of the peakedness of a probability density, usually relative to the normal density. A density that is peaked considerably more than the normal density is referred to as a *leptokurtic density*, while one that is relatively flat-topped relative to normal is referred to as a *platykurtic density*. A density that is about the same as the normal density is referred to as *mesokurtic*. Examples of such densities are shown in Figure A.8.

One measure of the kurtosis of a probability density is given by

$$\kappa = \frac{\mu_4}{\sigma^4} \tag{A.24}$$

where μ_4 is the fourth moment about the mean of the random variable, as defined previously, and σ is the aforementioned standard deviation.

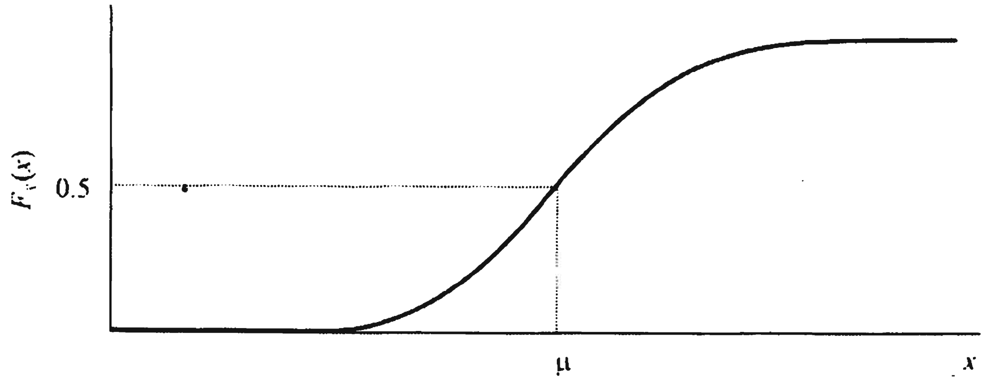


Figure A.7 Gaussian, or normal, continuous probability distribution function.

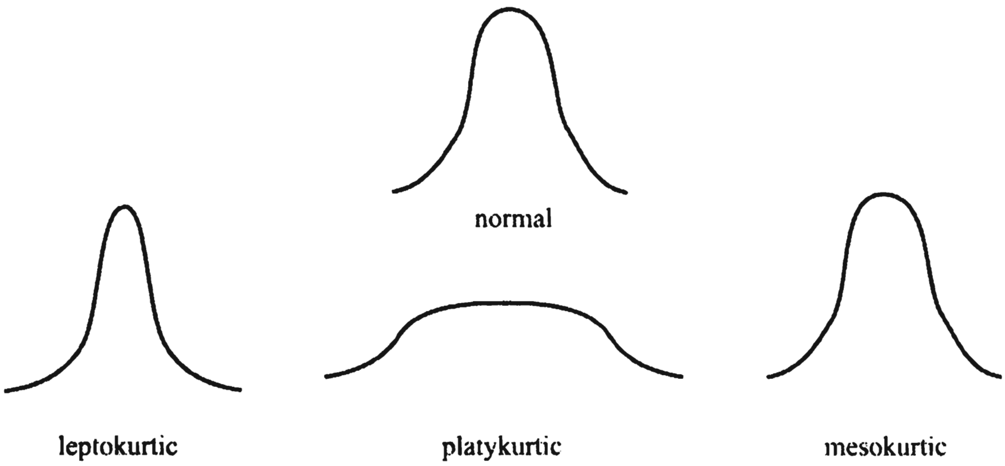


Figure A.8 Non-normal probability densities.

A.6 Skewness

The *skewness* of a probability density is a measure of the asymmetry of the density. Herein it is denoted by ψ. If the tail to the right compared to the left of the mean is larger and/or longer, then the density is said to be skewed to the right or have *positive skewness*. This is similar for the left tail. The normal definition of the skewness is given by

$$\psi = \frac{\mathcal{E}^3\{x-\mu\}}{\sigma^3} \tag{A.25}$$

A.7 Useful Characteristics of Probabilities, Density Functions, and Distribution Functions

Some useful properties of probabilities and their functions are presented in this section. The probability of the certain event is 1. There can be no probability larger than this. Thus, $P \leq 1$. Furthermore, the probability of the occurrence of an event cannot be less than zero; thus, $P \geq 0$.

The probability distribution function $P(x)$ for a random variable x is monotonically increasing. It is zero for some value of x and increases to 1 for some larger value of x. It cannot be larger than one or less than zero.

Since

$$F_X(x) = \int_{-\infty}^{x} f_X(\xi)d\xi \tag{A.26}$$

then $F_X(x_1) \leq F_X(x_2)$ whenever $x_1 \leq x_2$. Furthermore, the total area under $f_X(x) = 1$.

A.8 Concluding Remarks

This appendix provides the reader with an introduction to the topics in probability and random process theory necessary to comprehend the remainder of this book. Statistical processes is a large topic with a huge body of literature associated with it. It is worthy of study in its own right.

Communication systems, and the EW systems designed to counter them, process signals that are statistical in nature. This is due in part to the random noise that is invariably added to such signals as they are transmitted and processed. Some signals, however, exhibit random characteristics in and of themselves. The randomness is built into the communication signals by design. A simple example of this is encrypted signals, but there are many other examples. Therefore, an understanding of the fundamentals of statistical processes is important for a thorough understanding of the design of EW systems.

References

[1] Kosko, B., *Fuzzy Thinking*, New York: Hyperion, 1993, pp. 44–64.

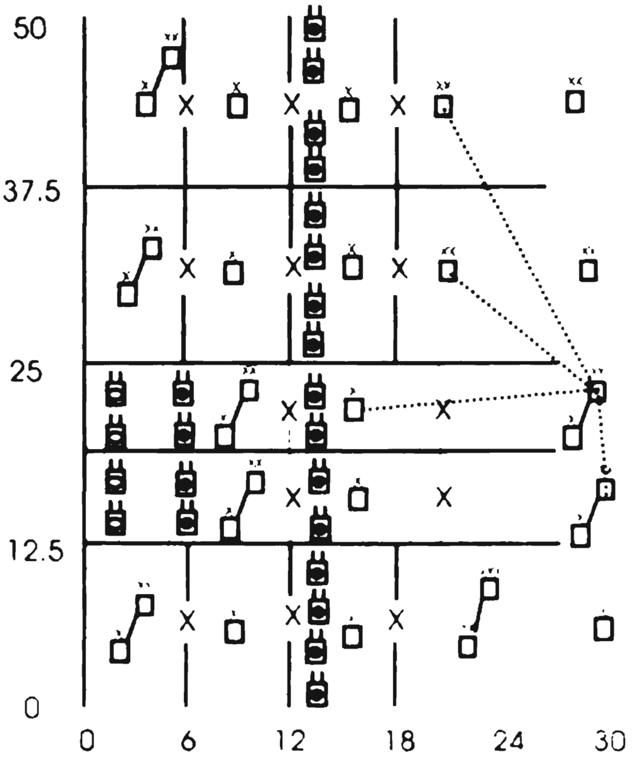
Appendix B

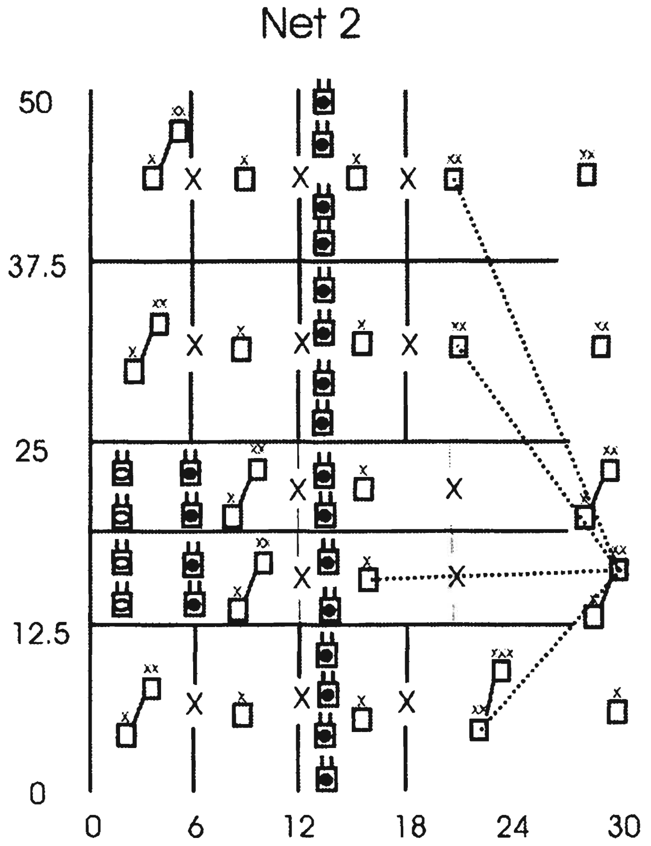
Simulated Networks

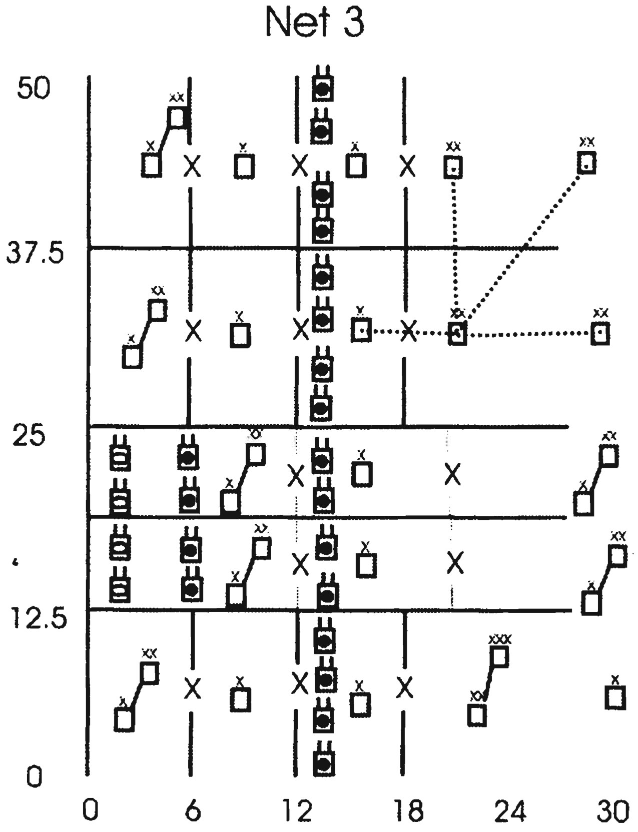
B.1 Introduction

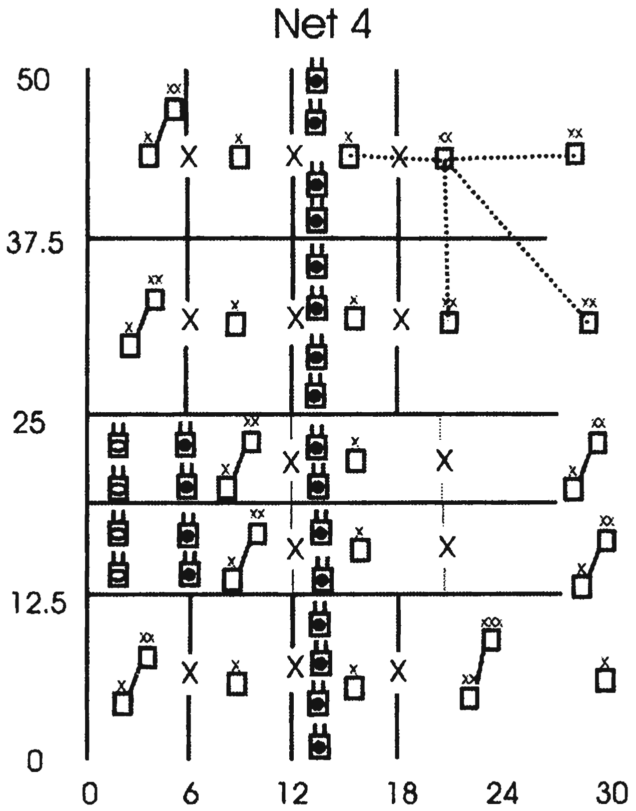
The specific nets used in the simulations described in the text are shown in this appendix. In those cases where there were only these 19 nets, no other nets were simulated. In those cases when more than 19 nets were simulated the additional nets were added to the region shown at random, but at tactically significant ranges.

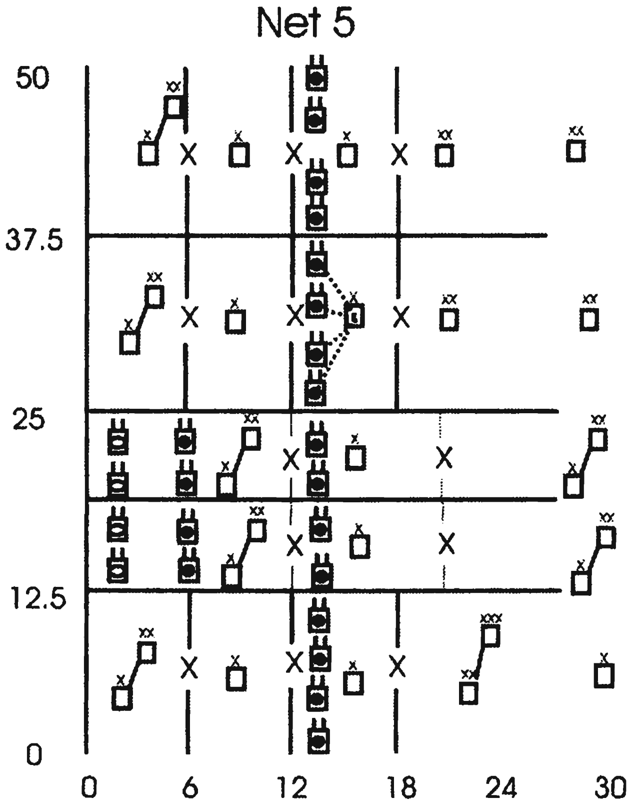
Net 1

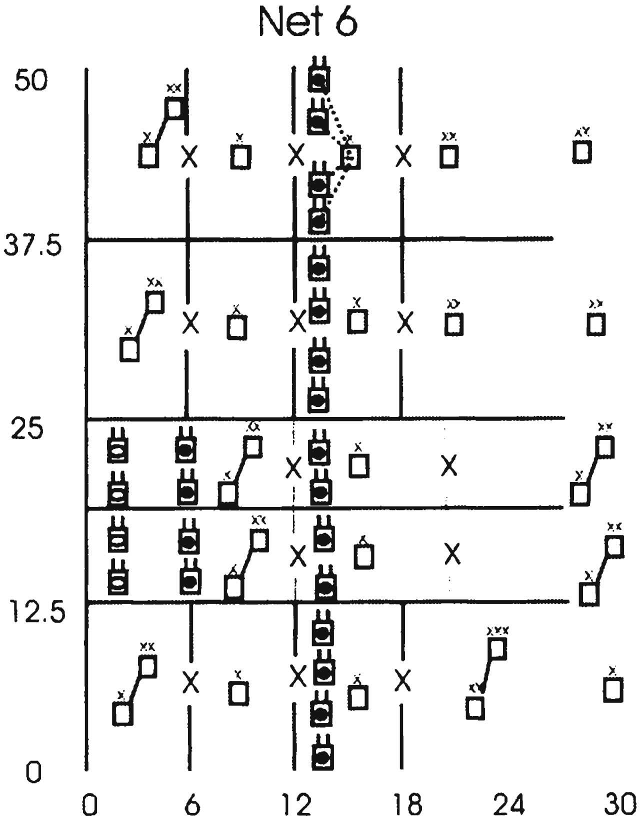


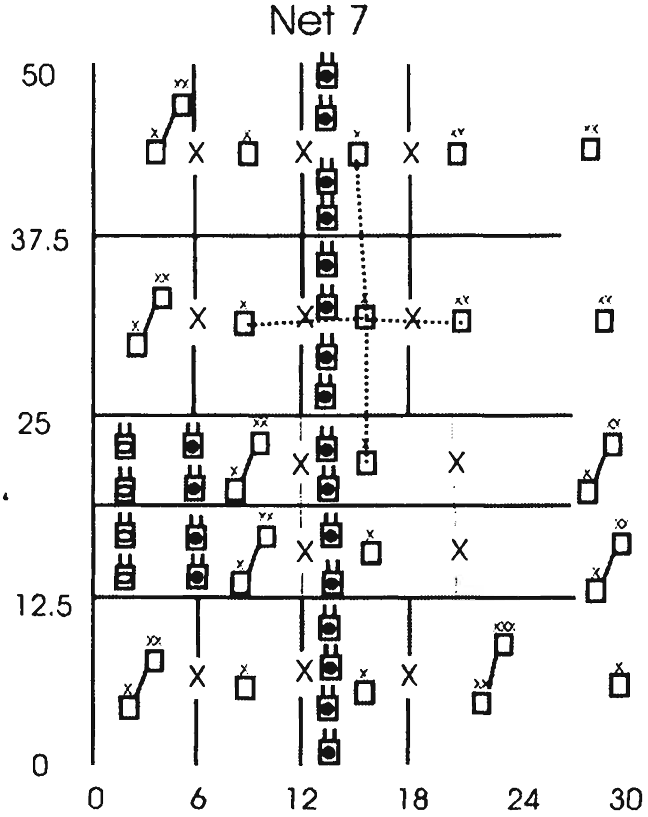

Net 2
50
37.5
25
12.5
0
0
6
12
18
24
30

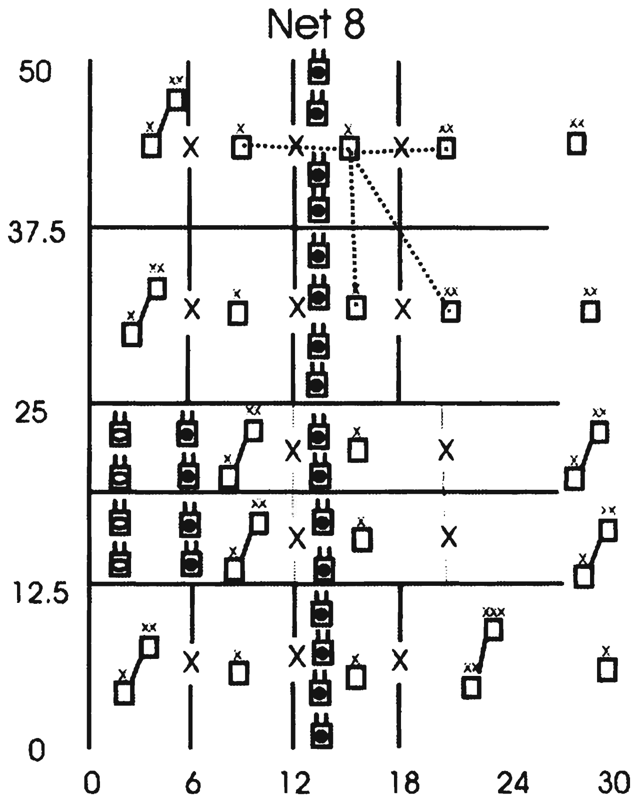

Net 3
50
37.5
25
12.5
0
0
6
12
18
24
30

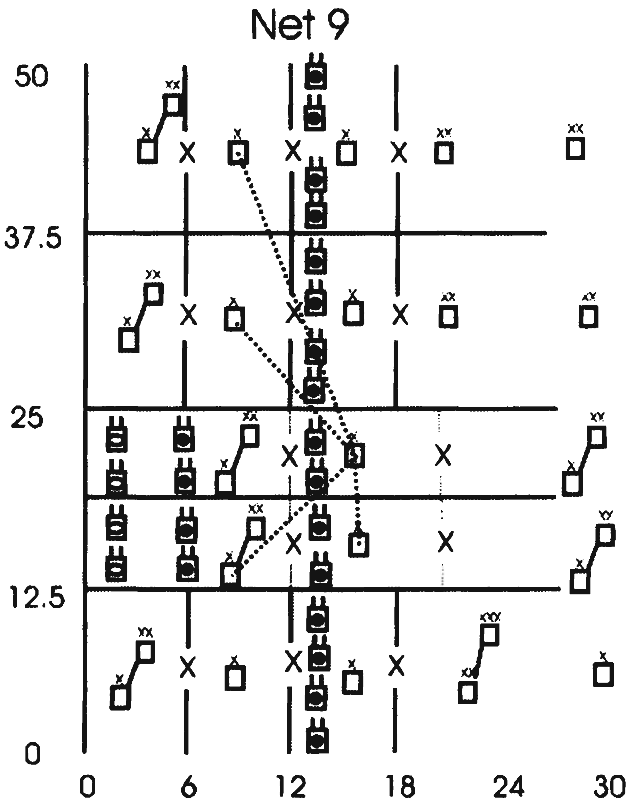

Net 4
50
37.5
25
12.5
0
0
6
12
18
24
30

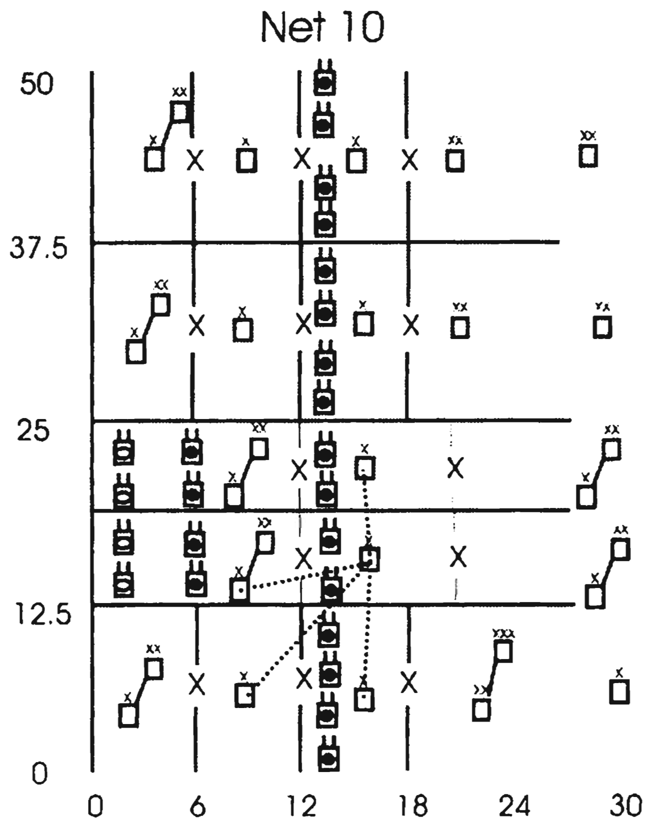

Net 5
50
37.5
25
12.5
0
0
6
12
18
24
30

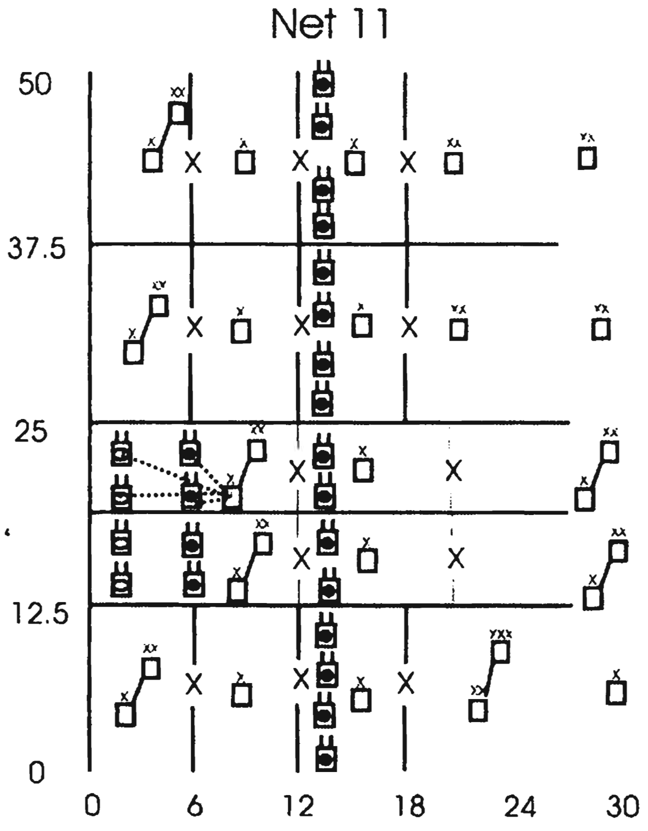

Net 6
50
37.5
25
12.5
0
0
6
12
18
24
30

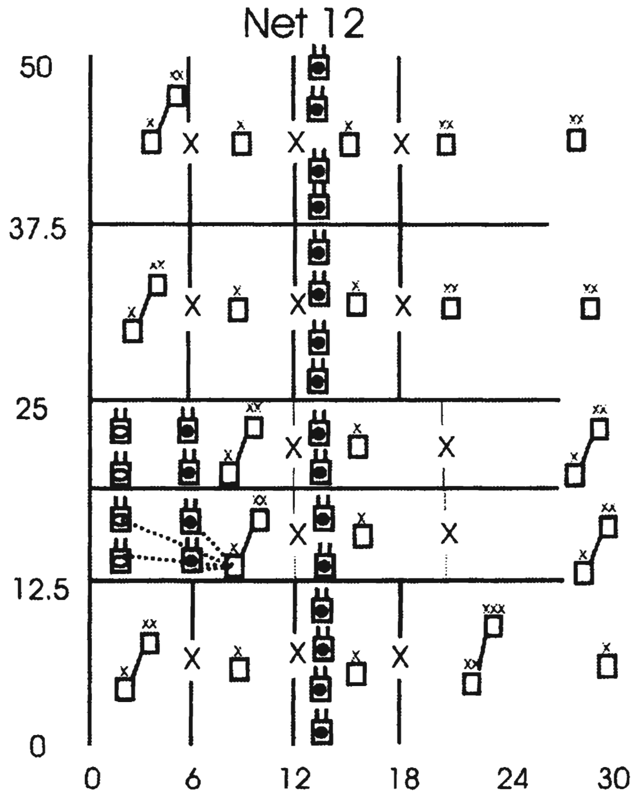

Net 7
50
37.5
25
12.5
0
0
6
12
18
24
30

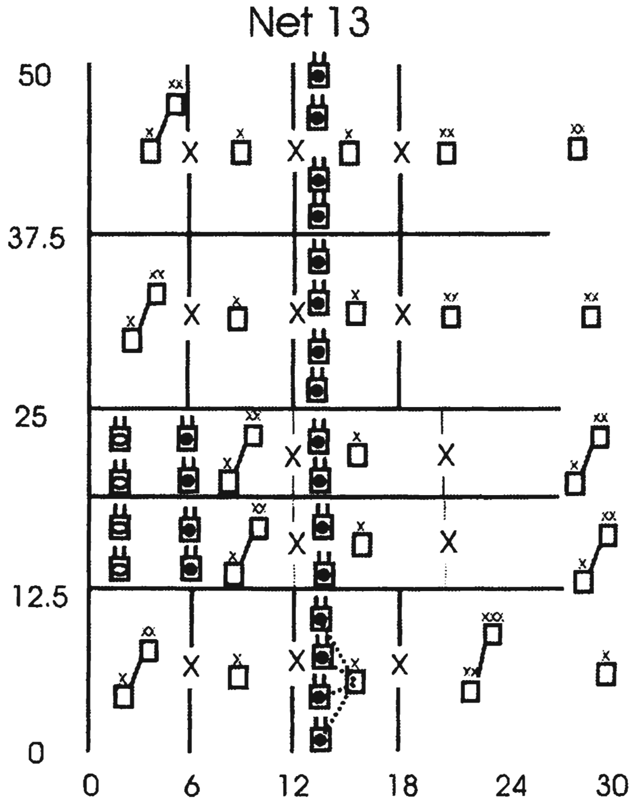

Net 8
50
37.5
25
12.5
0
0
6
12
18
24
30


Net 9
50
37.5
25
12.5
0
0
6
12
18
24
30

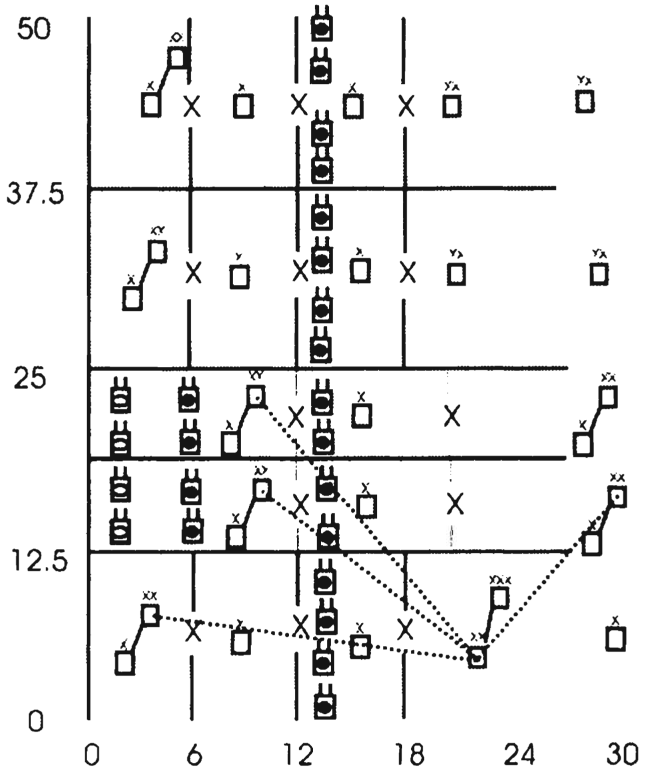

Net 10
50
37.5
25
12.5
0
0
6
12
18
24
30


Net 11
50
37.5
25
12.5
0
0
6
12
18
24
30

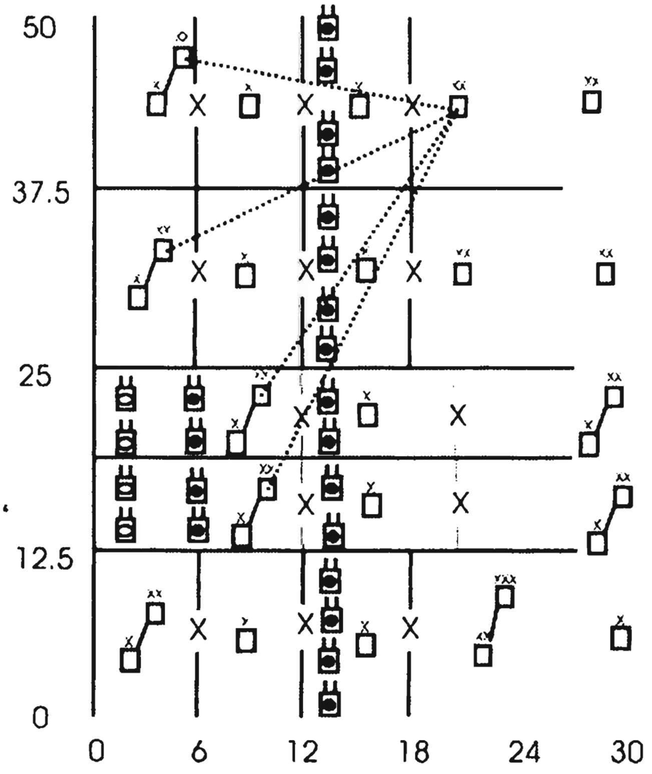

Net 12
50
37.5
25
12.5
0
0
6
12
18
24
30

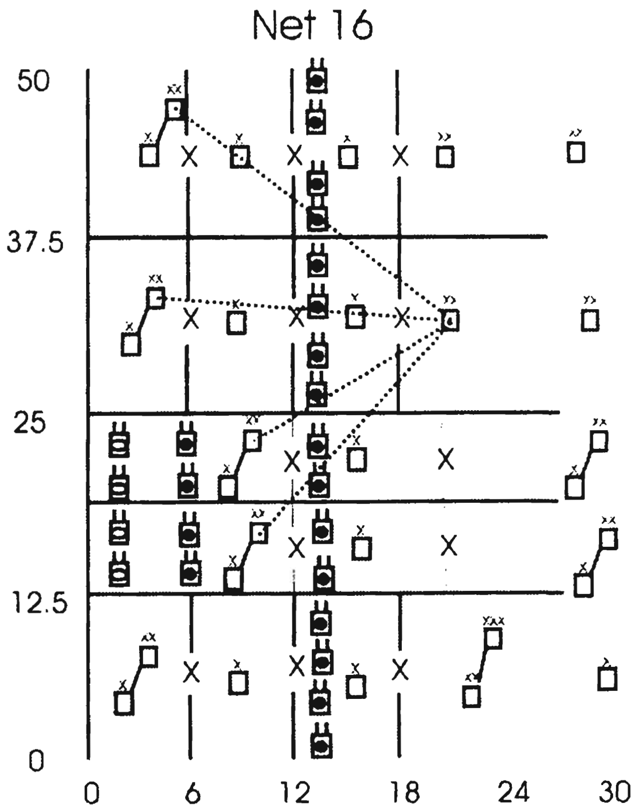

Net 13
50
37.5
25
12.5
0
0
6
12
18
24
30

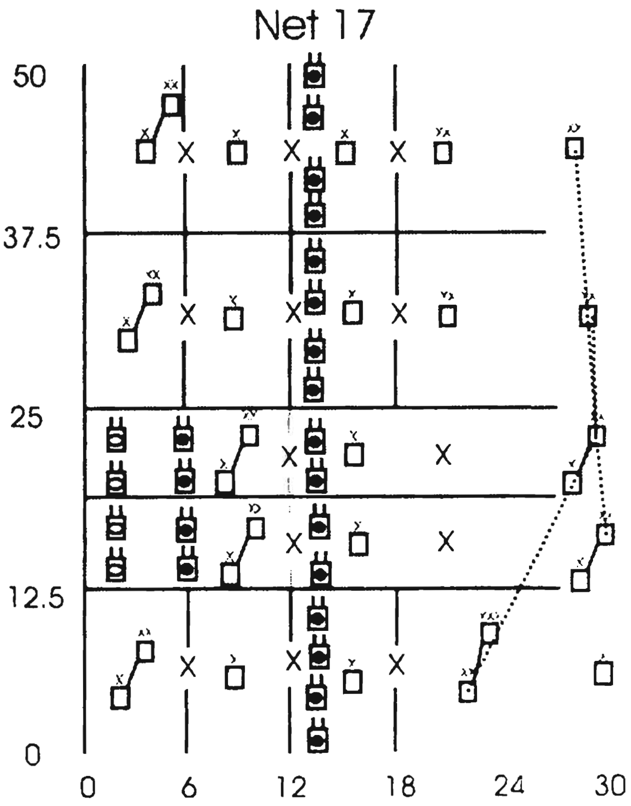
Net 14

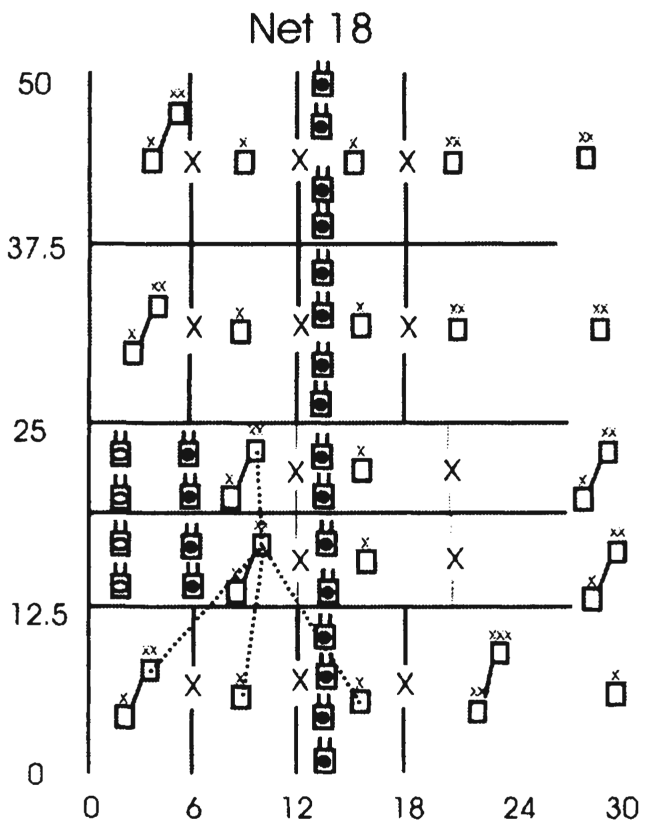


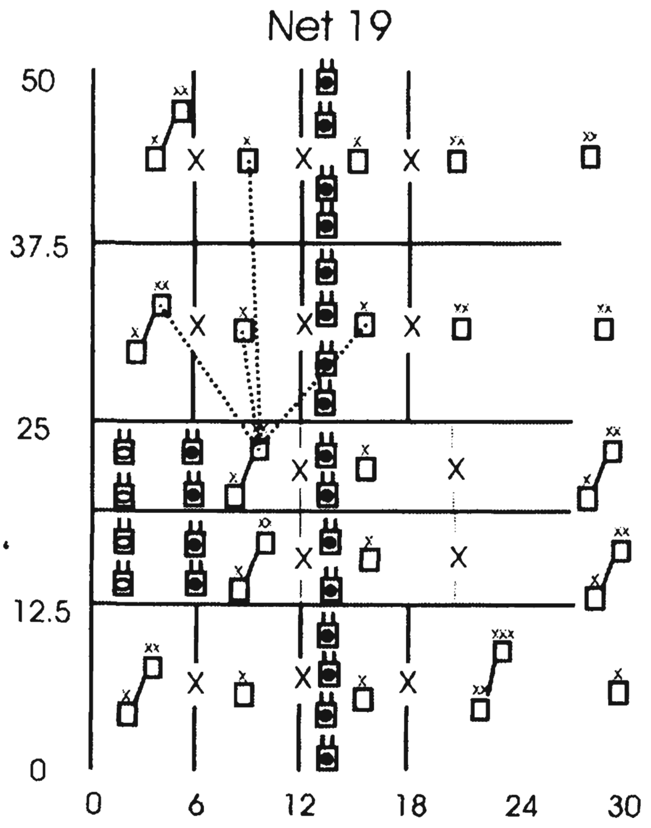
Net 15




Net 16
50
37.5
25
12.5
0
0
6
12
18
24
30


Net 17
50
37.5
25
12.5
0
0
6
12
18
24
30


Net 18
50
37.5
25
12.5
0
0
6
12
18
24
30


Net 19
50
37.5
25
12.5
0
0
6
12
18
24
30

Appendix C

System Engineering

C.1 Introduction

This appendix introduces the notion of system engineering, in the context of treating the system as a whole. "System" in this case is more than just the hardware manifestation normally thought of when the term is used. This will be discussed more fully in Section C.2. Suffice it to say that the term system includes all aspects of designing, building, and sustaining the system.

The design of any system, including EW systems, involves many disciplines. Introductions to some of the more important of these disciplines, including system reliability, fault tolerance, environmental concerns, EMI and *EM compatibility* (EMC), human factors, and safety, are presented in this chapter. Each of these topics could require a book in themselves to cover adequately, so what is included here are introductions to acquaint the reader with the topics at a very high level—sufficient to be able to address the topics at a cursory level if required.

C.2 System Engineering

Kayton defines system engineering as consisting of the following elements [1]:

1. Translate an operational need into a system;
2. Integrate all technical disciplines;
3. Ensure functional and physical interfaces;
4. Identify and abate risks;
5. Verify that the design meets the need.

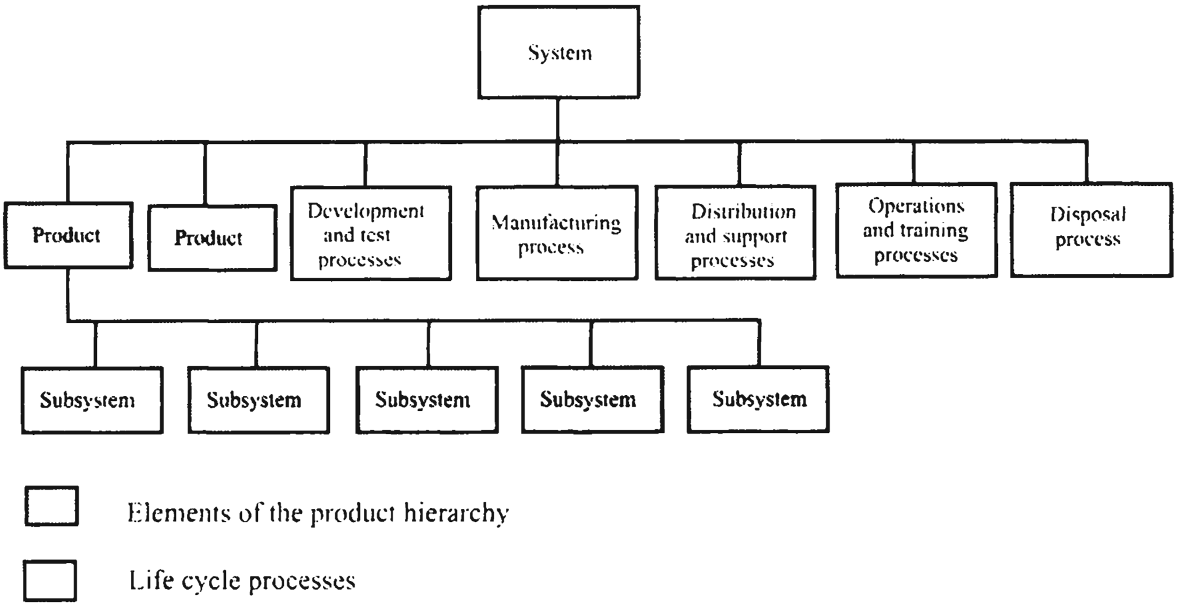

System
Product
Product
Development and test processes
Manufacturing process
Distribution and support processes
Operations and training processes
Disposal process
Subsystem
Subsystem
Subsystem
Subsystem
Subsystem
Elements of the product hierarchy
Life cycle processes

Figure C.1 Functional areas that need to be considered in the design and use of any system. (*Source:* [2]. © IEEE 1998. Reprinted with permission.)

The aspects involved with every system are shown in Figure C.1 [2]. Note that only two of the first level blocks refer to the system itself—the hardware, for example. The remainder refers to auxiliary activities affecting the life cycle of the system. The life cycle of a system starts with the research and development phase and ends with the disposal phase, where it is removed from useful service. Manufacturing, sales, distribution, training, and operation are all phases between these. The system engineering process must address all these phases and how they interrelate with one another for any useful system.

The life cycle of a system is notionally illustrated in Figure C.2 from IEEE Std. 1220–1998. It begins with the system definition, based on satisfying some requirement. The life cycle proceeds into subsystem definition, which is followed by system production and customer support stages, which overlap. The subsystem definition consists of the three steps shown in Figure C.2: preliminary design followed by detailed design, which is then followed by the *first article integration and test* (FAIT) phase.

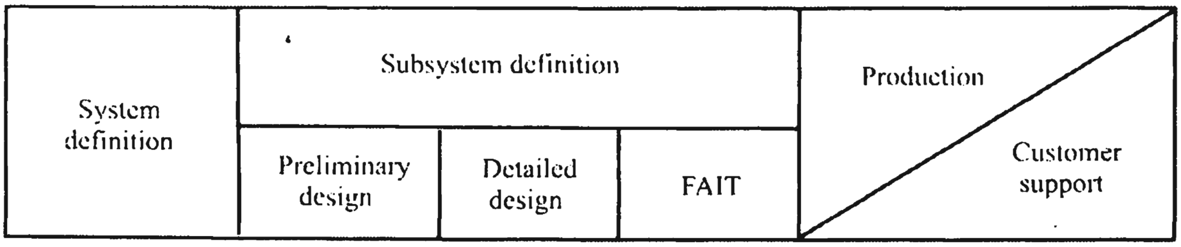


Figure C.2 Life cycle of a system. (*Source:* [2]. © IEEE 1998. Reprinted with permission.)

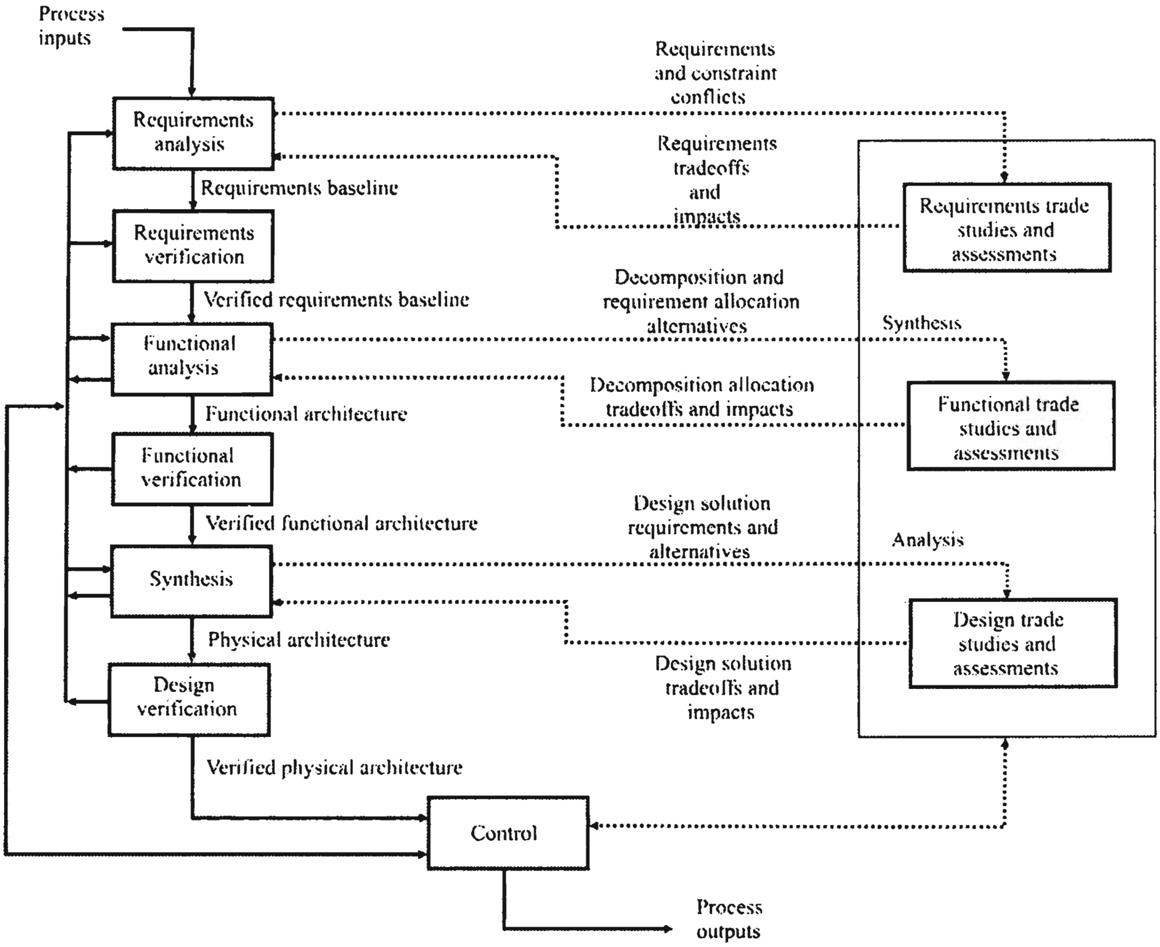

Process inputs
Requirements analysis
Requirements baseline
Requirements verification
Verified requirements baseline
Functional analysis
Functional architecture
Functional verification
Verified functional architecture
Synthesis
Physical architecture
Design verification
Verified physical architecture
Control
Process outputs
Requirements and constraint conflicts
Requirements tradeoffs and impacts
Decomposition and requirement allocation alternatives
Decomposition allocation tradeoffs and impacts
Design solution requirements and alternatives
Design solution tradeoffs and impacts
Requirements trade studies and assessments
Synthesis
Functional trade studies and assessments
Analysis
Design trade studies and assessments

Figure C.3 Flow diagram illustrating the system engineering process. (*Source:* [2]. © IEEE 1998. Reprinted with permission.)

The *system engineering process* (SEP), according to IEEE Standard 1220–1998, follows the flow diagram shown in Figure C.3. The steps indicated in this process are repeated for each and every action associated with system engineering—all the way from designing the overall system itself, to analyzing incremental changes to the system.

The first step is to analyze the requirements that the system is to address. Tradeoff analyses are performed on the requirements, as there are frequently conflicts in these requirements—they cannot all be met. A requirements baseline is thus established. These requirements are then verified and validated, usually with a customer or user of the system. The functions of the system are then defined with further trade studies and with subsequent verification and validation. Functions are allocated to subsystems, to include software versus hardware allocations. Based on these allocations, the synthesis process begins to actually design and build the physical system, again to include hardware and software. These designs and

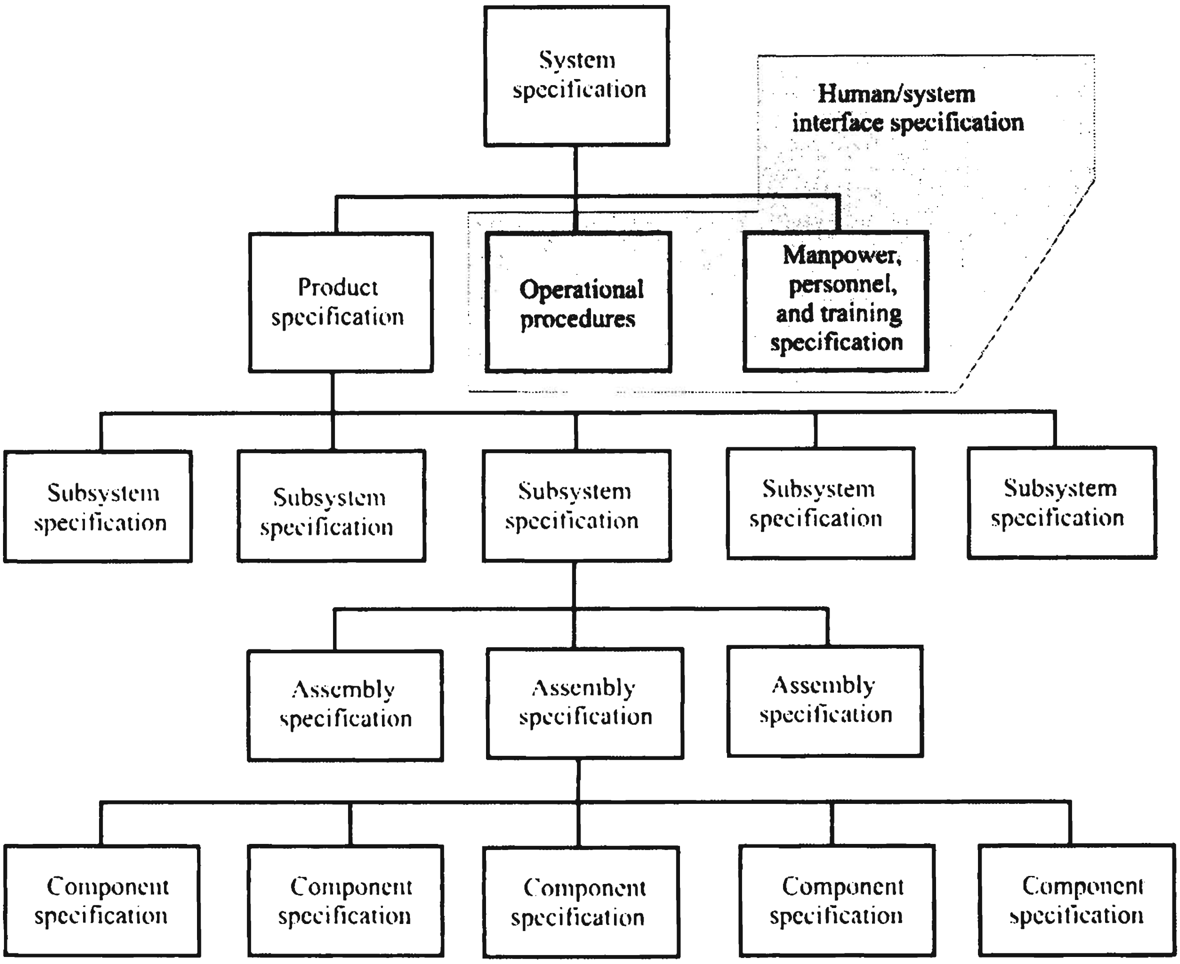


Figure C.4 The specification tree. (*Source:* [2]. © IEEE 1998. Reprinted with permission.)

physical manifestations are then verified, usually to include substantial testing. Once the functions are verified in the physical system, the process is either iterated for design refinement, where feedback can go to any of the previous steps, or the process terminates with a completed system.

The early stages in the development and deployment of any system involve system design, which primarily consists of performing tradeoff analyses to ascertain the best system design. Again, Kayton indicates that these tradeoffs almost always boil down to deciding between the characteristics of "performance, reliability and availability, human convenience, and cost." System engineering considerations described below will be described following these basic categories.

Perhaps the most important output of the initial system engineering process is the system specification. This specification, according to IEEE Standard 1220–1998, consists of the elements shown in Figure C.4. The overall system specification consists of the three parts shown: the product specification, the operational procedures used with the system, and the human aspects consisting of the manpower, personnel, and training specification. The product specification is subdivided as shown in Figure C.4, eventually consisting of specifications at the component level.

With the advent of very highly integrated semiconductor chips, the system engineering process is changing. There is and will be a great deal of emphasis on simulation and modeling in order to establish the system specification. Codesign is the cornerstone to this new philosophy, where all aspects of system design occur simultaneously.

C.2.1 Performance Characteristics

The other chapters of this book discuss the performance characteristics more or less unique to communication EW systems. This chapter discusses some of the system design considerations that apply to systems in general, not just communication EW systems. They thus deal with performance characteristics at the whole system level, as opposed to specific subsystems or sub-capabilities.

C.2.2 Environmental Characteristics

Shown in Table C.1 are the environmental specifications for computer and electronic equipment from IEEE Standard 1156.2–1996 for the case when the equipment is nonoperating. Environmental conditions for tactical military equipment are typically more severe than those for commercial equipment, although most of the areas of concern are the same. The conditions shown in Table C.1 are commercial specifications so they will apply to those communication EW systems that are built for this environment. PL1 refers to equipment subject to moderate amounts of the condition in question. PL2 refers to the same equipment subjected to minimal amounts. If the unit passes the test, then it can be expected to be able to survive the environmental conditions specified. The specific conditions are shown in the left column of Table C.1. Table C.1 contains the conditions for testing for the condition indicated.

Low- and high-temperature environments are concerned with the shipment and storage conditions, after which the equipment will still work. Frequently equipment is shipped and/or stored at temperature extremes because the shipping/storage containers are not temperature controlled. The test is of such duration to simulate a cold or hot soaking, so that the equipment is settled at that temperature.

Electronic equipment can be put into service at a moment's notice from extremely cold conditions. The purpose of the thermal shock test is to verify that the equipment will operate after such a transition. Note that the test conditions go from –40°C to +65°C in a period of only 5 minutes.

Table C.1 Nonoperating Conditions

Test Parameter	**Test Publication**	**Severity or Conditions**	**Requirements**	
			PL1	PL2
Low temperature	IEC 68-2-1 (1990) Test Ab	24 hours	–40°C	–60°C
High temperature	IEC 68-2-2 (1974) Test Bb	24 hours	+65°C	+55°C
Thermal shock	IEC 68-2-14 (1984) Test Na	< 5 min	–40°C to +65°C	N/A
Humidity	IEC 68-2-30 (1980) Test Dd	6 cycles	25°–55°C 92 ± 3% RH	25°–55°C 92 ± 3% RH
Free-fall packaged	IEC 68-2-32 (1975)	Corners and faces	Combinations	Faces only
Handling unpackaged	IEC 68-2-32 (1975)	Corners and faces	Combinations	Combinations
Vibration	IEC 68-2-6 (1975)	5–200/300 Hz (see Figure C.5)		
Shock	IEC 68-2-27 (1987)		30g/20 ms	15g/20 ms
Mixed flowing gas	ASMT 8-827-92		14 days	7 days
Fungus	IEC 68-2-10 (1988) Test J		X	N/A
Flammability	System test		ANSE T1319-1995	UL 94-1991
Electrostatic discharge	IEC 801-2 (1991)	15 kV air discharge	X	X

Source: [3].

Some military equipment must be designed to operate in high-humidity environments such as Central American jungles. Thus, the requirement for military equipment can be more severe than commercial equipment because of this. The humidity conditions in the table are relatively benign.

The free-fall tests are designed to evaluate whether the equipment can be dropped under normal handling conditions. Such conditions could be as simple as removing the equipment from the shipping carton and installing it in a rack. The distance the equipment must fall depends on its weight.

The amount of shock and vibration a system must tolerate is dictated to a large extent by the operational environment in which that system is used. For example, a radio intended to operate in a tank must be designed to tolerate the mechanical environment of that tank, which can be quite harsh. On the other hand, a desktop computer designed to never leave the desktop (except for repairs) survives in a benign environment and shipping and handling procedures would probably dictate the amount of shock and vibration it must tolerate.

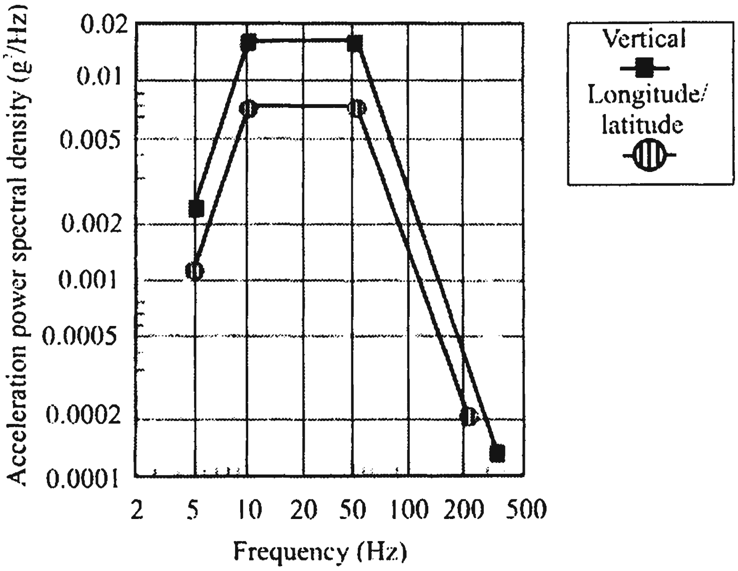


Figure C.5 Vibration characteristics when a system is nonoperating. (*Source:* [3]. © IEEE 1996. Reprinted with permission.)

Shock and vibration are two fundamentally different parameters. Shock refers to the relatively infrequent sharp impulse of mechanical energy coupled into a system from the likes of dropping an electronic enclosure onto a concrete floor. Values of such shock can reach 50g or more, where 1g = the force of gravity. Vibration, on the other hand, refers to the normally repetitive mechanical oscillation to which a device can be subject. An example of this might be the vibration induced into a system by the rotation of helicopter blades.

Mechanical stability is normally a design requirement whether the EW system is strategically deployed, never to be moved, or tactically designed, intended to be moved frequently. The latter can be over benign traveling conditions or rough, battlefield conditions.

The spectrum of the vibration specified in IEEE Standard 1156.2–1996 is shown in Figure C.5 for nonoperating equipment. The highest acceleration, and therefore the highest vibration amplitude, is in the frequency range of 10–50 Hz.

Equipment deployed to humid regions frequently encounter conditions that support the growth of fungi. The fungus test will verify that the equipment does not degrade under these conditions. It should be noted that equipment subject to this testing is frequently not usable afterward, for the most part, because of the debilitating effects of the fungus growth.

Flammability is of concern especially for electronic equipment. Such equipment can be expected to operate in hot conditions, and if the components were prone to burn, significant safety issues arise, not to mention whether the equipment will continue to operate. In a military setting, equipment can be exposed to conditions that could promote burning, such as when under hostile fire.

Table C.2 Operating Conditions

Test Parameter	Test Publication	Severity or Conditions	Requirements	
			PL1	PL2
Low temperature	IEC 68-2-1 (1990) Test Ad	4 hours	+5°C	+10°C
High temperature	IEC 68-2-2 (1974) Test Bd	4 hours	+50°C	+35°C
Sine vibration	IEC 68-2-6 (1995) Test Fc	5–100 Hz	0.5 mm/0.1g	0.5 mm/0.1g
Shock	IEC 68-2-27 (1987)	3 pulses	15g/11 ms	10g/11 ms
Low air pressure	IEC 68-2-13 (1983)	6 hours	10.5–C.5 Pa	10.5–7.85 Pa
Electrostatic discharge	IEC 801-2 (1991)	15 kV/150 pF	X	X
Electromagnetic induction	FCC rules part 15 IEC 801-3 (1984)	Meet	X	X
Immunity	IEC 801-4 (1988)	dc power	X	X
Immunity	IEC 801-4 (1988)	ac power	X	X
Immunity	IEC 801-4 (1988)	I/O ports	X	X
Earthquake	IEC 68-2-57 (1989)		X	N/A

Source: [3].

The mixed flowing gas test is a simulation of the effects of being exposed to gases to evaluate metal corrosion. The gaseous test exposes metal components to corrosive gases for a period of time during which thermal cycling is performed. Typical test conditions consist of 25 or more thermal cycles in gaseous mixtures consisting of the low parts per billion of water, nitrogen oxide, and chlorine.

Electrostatic discharge testing determines whether the equipment will withstand the typical electrical discharge that occurs when one walks across a carpet and touches something metallic. Electronic equipment designed for human interaction, including operation and/or repair, will encounter such discharge.

Salt fog testing is not included in Table C.1, but is frequently a requirement of military equipment. It is not so much of a concern to commercial equipment. Military deployments typically involve shipment of electronic equipment, either the system itself, or components of it, say, spare parts, by sea. This environment frequently consists of salt in very humid air.

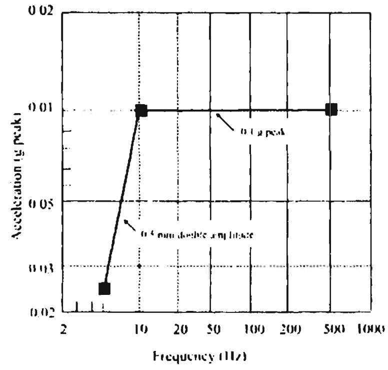


Figure C.6 Operating vibration characteristics. (*Source:* [3]. © IEEE 1996. Reprinted with permission.)

Similar specifications for electronic equipment under operating conditions are shown in Table C.2. A few of the conditions are different, there are a few additional tests, and some of the tests have been deleted.

The vibration specification is somewhat different for operating versus nonoperating equipment. The operating spectrum is shown in Figure C.6. Here the acceleration peak amplitude is about the same as nonoperating, but the spectrum extent of the vibration is extended from 50–500 Hz.

Low-air pressure tests are included to evaluate the effects of altitude on the operation of the equipment. As the altitude is increased, the air is less dense and there is less air pressure since air pressure is caused by the weight of the air above it and there is less air the higher the altitude. Components can sometimes become inoperable when assembled at one air pressure and operated at a lower (or higher) pressure. Airborne EW equipment is not always located in pressurized space on the aircraft. It therefore can be subjected to very low pressure, as some aircraft fly at significant altitudes. EW equipment that is intended to be used in such applications must be specifically designed for low-pressure operation.

EW systems, by their very nature, are susceptible to unintentional wandering signals. Careful design practices must be followed in order to avoid interfering signals from being generated from within the system itself. Such signals are referred to as EMI.

EW systems are particularly susceptible to EMI because of the very sensitive receivers used for ES. These receivers frequently have wide instantaneous bandwidths, and if not instantaneous, then total bandwidths. Even narrowband receivers have filters at their input that can be as wide as half an octave.

One of the design parameters for EW systems is the system sensitivity. The more sensitive the system, the better. However, more sensitivity implies more susceptibility to unwanted noisy EMI signals. Shielding is almost always required at the box level, and frequently at the circuit-board level. Copper cabling that connects the boxes together must be shielded. Fiber-optic interconnects hold promise for alleviating some of the EMI problems when interconnecting boxes together as well as within electronic enclosures themselves.

The high-speed clocks that are designed to approximate square waves in computer and other high-speed digital circuits are a prime culprit for generating EMI. The frequency spectrum of a square wave theoretically extends to infinity in both directions around the fundamental frequency. Some of these components can be significantly large as well. Therefore, careful shielding of such clock signals is required. In fact, in the design of digital circuits, if a slower clock can be tolerated, it should be used. High-speed signal and clock cables should be carefully routed along the chassis, if appropriate, and cross other signal lines at right angles to minimize mutual coupling, for example.

EMC refers to the ability to operate in the presence of other systems without them interfering with the operation of the EW system or the EW system interfering with them. Such systems are said to be compatible with one another. EMC testing is referred to as *immunity* in Table C.2. EMC is substantially more difficult in EW systems that employ EA because of the high power signals normally associated with EA. Tactical communication EW systems are infrequently required to operate in earthquake conditions, so this specification is rarely applied.

C.2.3 Reliability and Availability

The ability of a system to operate when required is key. This is particularly true for systems discussed herein, where battles can be won or lost and lives saved or lost depending on whether a system operates correctly when required. Reliable systems do not emerge by accident. Conscious effort must be expended to increase the reliability of an otherwise unreliable system. This section introduces the notions and basic terminology associated with designing a system to be reliable.

C.2.3.1 Why Reliable Design Is Necessary

All real systems fail at some point. It is possible to design systems so that their tolerance to failure of components and subsystems can be improved, however. Such design always adds redundancy of some type, thus normally increasing such factors as the cost, the system size, and power consumption. Because of this, redundancy for only critical components, that is, those that are necessary to insure that some facet of the operation of the system is guaranteed to work when required, is added. An example of this might be the central server in a computer network. If its operation is critical for operation of the network, then a second server, performing the same function, could be added. In contrast, one of the work stations using the server on the network might be allowed to fail without causing the functioning of the network to fail totally.

As systems get more complex due to added components and functionality, their reliability declines. Reliability in this sense means the probability that the system will continue to function without failure. Failures can be categorized depending on their importance to the successful completion of the mission of the system. The engine in a tank is a critical subsystem that must continue to function for the tank to move, and normally movement is an important part of a tank's mission. The antenna used for command and control communications for the tank can probably fail and the tank remain useful for most of its intended functions. A failure that precludes a system from accomplishing its mission is called a *critical failure*. One that is not critical is called *noncritical*. The aforementioned tank engine may not be a critical failure if the mission of the tank does not call for it to move. A tank sitting at a road junction in an over watch mission does not need to move, so if its engine fails while it is sitting there, it is a noncritical failure for that mission.

Reliability of complex electromechanical systems, without designing in some degree of fault tolerance, is typically very low, requiring exorbitant amounts of repair actions and nonavailability of the system when it is required. *Built-in test equipment* (BITE) helps, but is intended to assist a technician with repair actions as opposed to continuing operation once a failure has occurred. Whereas the down-time of the system can be improved, and thus the ratio of operational time to total time is improved, the mean time between failure is not improved. The reliability of systems can be calculated, or at least estimated, based on the reliability of the individual components used in the systems.

One of the attractions of the evolving *microelectrical-mechanical* (MEM) technology is to design extensive redundancy into an otherwise nonredundant system. Microtechnologies such as these facilitate redundancy without unacceptably increasing system size.

C.2.3.2 Reliability

The time between system failures is a random variable with statistical properties. Let $R(t)$ denote the reliability function, which is the probability that the system does not fail at time t. It is frequently assumed that $R(t) = e^{-\lambda t}$ where λ is the failure rate. The amount of time between critical failures of a system is called the *mean time between failures* (MTBF). Then MTBF = $1/\lambda$. For complex electronic systems, typified by modern communication EW systems, the MTBF of systems designed without much fault tolerance could be on the order of 10 hours. The MTBF for the space shuttle, which has extensive fault tolerance built in, is on the order of weeks or months to ensure safe returns. If $\lambda = 1$ failure per 1,000 (1/1,000) hours, then $R = e^{-0.001t}$. This reliability function is shown in Figure C.7.

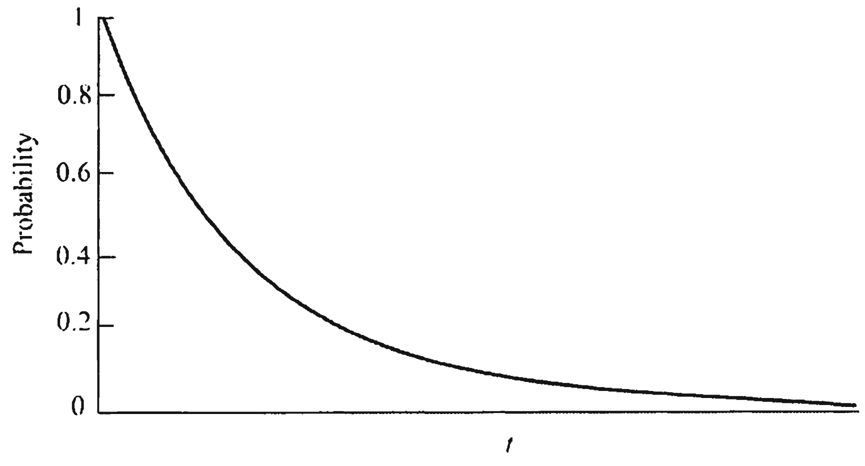


Figure C.7 A typical reliability function.

Once a system has failed, it would normally undergo some repair action by an appropriate technician. Systems can be designed to be easy to repair and they can be designed to be difficult to repair. An example is automobiles. Modern automobiles are diagnosed by computer, with repairs carried out by mechanics. Having an experienced mechanic carry out the diagnostics would make the cost exorbitant for the repairs. The average time to repair a system is called its *mean time to repair* (MTTR).

C.2.3.3 Tolerating System Failures

There are ways to improve the reliability of systems. It is always true, however, that these ways involve adding components to the system of some variety. Some of the ways of designing these tolerances to faults are introduced in this section.

If the success of accomplishing a mission is dependent on the reliability of two subsystems, then a reliability model of these two subsystems is as shown in Figure C.8. The system will fail if either of the two subsystems fails. Thus, the reliability of the system is given by

$$R = R_1 R_2 \tag{C.1}$$

because for the system to have survived to time t, subsystem 1 and subsystem 2 must have survived. If either of the subsystems surviving causes the mission to be

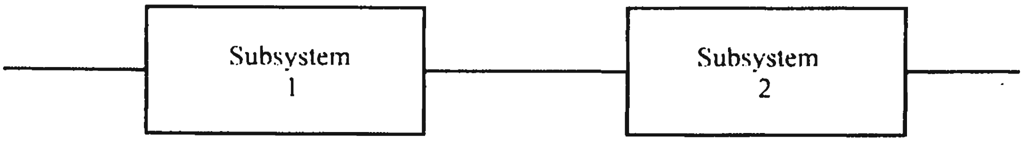


Figure C.8 Series reliability model.

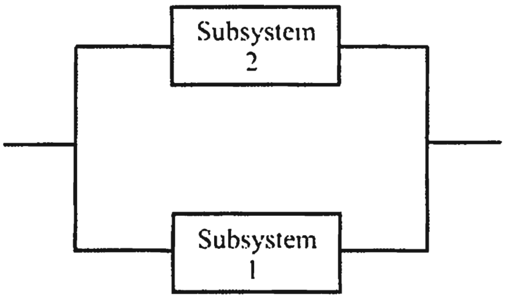


Figure C.9 Parallel reliability model.

accomplished, then the reliability model of the two subsystems is as shown in Figure C.9. Noting that P(failure) = 1 – $R(t)$, then the probability of the system failing is given by P(failure) = P_1(failure)P_2(failure). Thus, $R(t)$ = 1 – P(failure) = 1 – P_1(failure)P_2(failure) so

$$R(t) = 1 - [1 - R_1(t)][1 - R_2(t)] \tag{C.2}$$

The reliability model shown in Figure C.10 will fail only if three of the four subsystems fail. Its reliability can be shown, by derivations similar to that above, to be

$$R(t) = \{1 - [1 - R_1(t)][1 - R_2(t)]\}\,\{[1 - R_3(t)][1 - R_4(t)]\} \tag{C.3}$$

C.2.3.4 Failure Modes

The failure mode of components sometimes needs to be taken into consideration. The diode shown schematically in Figure C.11(a) has the ideal performance characteristic shown in Figure C.11(b). Whenever the voltage on the anode (left) side is greater than the voltage on the cathode (right) side, the diode is a short circuit and exhibits no resistance to current flow. On the other hand, if the voltage on the right is more positive than that on the left, the diode behaves as an open circuit, and no current flows. When two diodes are connected in series as shown in

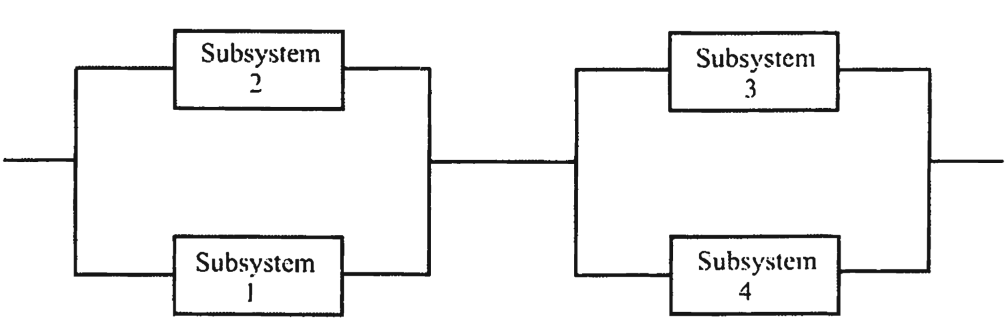


Figure C.10 Serial/parallel reliability model.

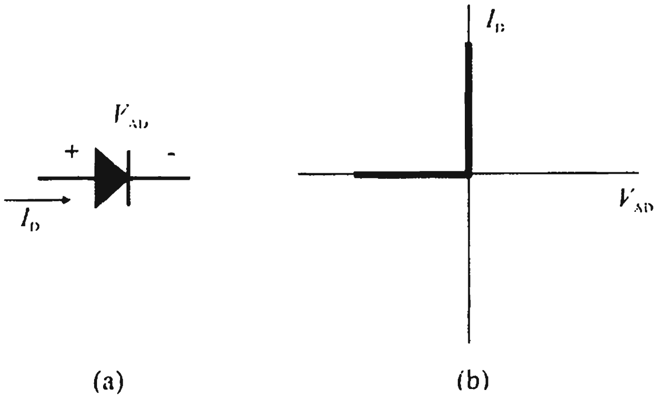


Figure C.11 (a) A diode and (b) its (ideal) characteristics.

Figure C.12, and the dominant failure mode of the diode is shorting, then failure of one of the diodes does not cause the circuit to fail. Thus, the probability of failure of the circuit is given by

$$P = P(\text{Failure of diode 1 and diode 2}) \tag{C.4}$$

If it can be assumed that the failure of one of these diodes is independent of the failure of the other (a significant assumption, especially for integrated circuits), then

$$P = P(\text{Failure of diode 1})P(\text{Failure of diode 2}) \tag{C.5}$$

and the reliability is given by

$$\begin{aligned} R &= 1 - P \\ &= 1 - P(\text{Failure of diode 1})P(\text{Failure of diode 2}) \end{aligned} \tag{C.6}$$

On the other hand, if the dominant failure mode of the diode in Figure C.11 is to open, then the configuration shown in Figure C.12 does not improve the reliability of the circuit. The circuit shown in Figure C.13, however, deals with this case, decreasing the overall failure rate and thus increasing the system reliability. If the dominant failure mode of the diodes shown in Figure C.11 is an open circuit failure, then for the circuit shown in Figure C.12, the probability of failure of the circuit is the same as the probability of failure of a single diode, and the overall

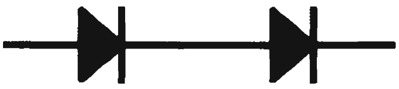


Figure C.12 When the dominant failure mode of the diodes is shorting, connecting two diodes in series will ensure the circuit will continue to operate.

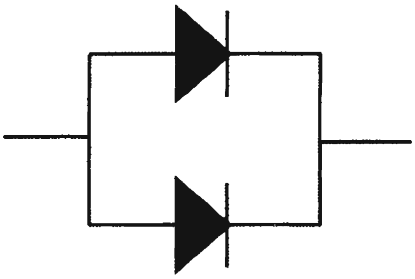


Figure C.13 Redundant design of a diode arrangement when the dominant failure mode is an open circuit.

reliability has not been improved by adding the second diode. This is because the probability of failure of the circuit is given by

$$\begin{aligned} P &= P(\text{Failure of diode 1 or diode 2}) \\ &= P_1 + P_2 - P_1 P_2 \end{aligned} \tag{C.7}$$

If the dominant failure mode is not known, or is absent, then the circuit shown in Figure C.14 will handle either when a diode fails by shorting or opening. If any one of the diodes fails, the circuit will continue to operate as designed. For some failure modes, the circuit will continue to operate even if up to three of the diodes fail. For example, if either of the two diodes on the left side shorts and if one of the diodes on the right open-circuits, then the circuit will continue to operate. Clearly this can get to be a fairly expensive process, increasing the component count, in this case, by a factor of four to accomplish the same function.

These examples demonstrate how fault tolerance can be built into electronic circuits if the functioning of a diode is the design component. To show how it can be applied to other types of components, consider the *field effect transistor* (FET) shown in Figure C.15. The transfer characteristic for this device is shown in Figure C.15(b). Properly integrated into a circuit, of course, as the voltage between the gate (G) and the source (S) is increased above zero, the channel between the drain (D) and source starts to close and conduct current. The gate settles at some voltage, saturating the device and the drain is essentially at the same voltage as the source.

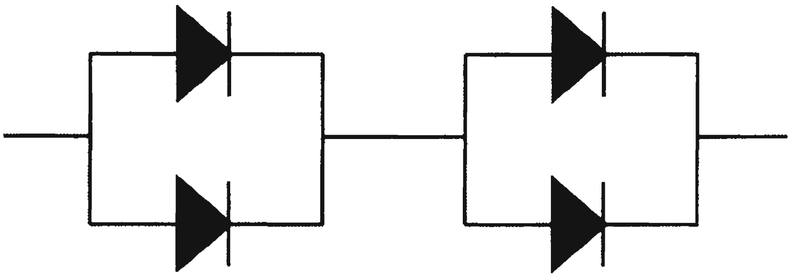
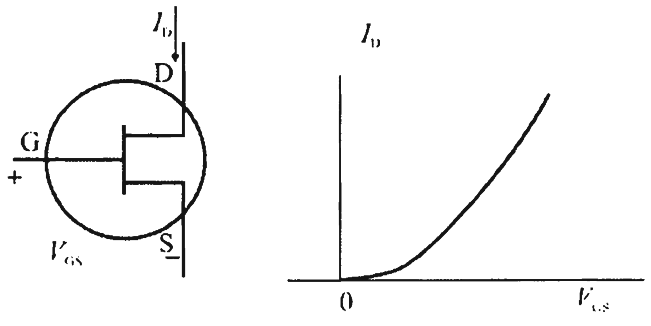


Figure C.14 Diode arrangement when the dominant failure mode is not known. Open or short failure of any one diode (as well as other selected failures) and the circuit will continue to operate as a diode.



(a) (b)

Figure C.15 (a) Schematic symbol for an FET and (b) the transfer characteristic of the FET.

If the dominant failure mode of the transistor shown in Figure C.15 is an open channel, then connecting two of the transistors in parallel as shown in Figure C.16 will provide additional reliability, since if one fails, the circuit will continue to function. In normal circuit operation, when

$$V_{GS} > 0, \text{ then } I_{D1} = I_{D2}, I_D > 0 \text{ and } V_o = 0$$
$$V_{GS} = 0, \text{ then } I_{D1} = I_{D2} = 0 \text{ and } I_D = 0,\ V_o > 0.$$

With a failed FET, say, FET$_1$, with an open channel then when

$$V_{GS1} > 0, I_{D1} = 0, I_{D2} > 0, I_D > 0 \text{ and } V_o = 0$$
$$V_{GS1} = 0, I_{D1} = 0, I_{D2} = 0, I_D = 0, \text{ and } V_o > 0.$$

Therefore, the circuit continues to operate as if there were no failure.

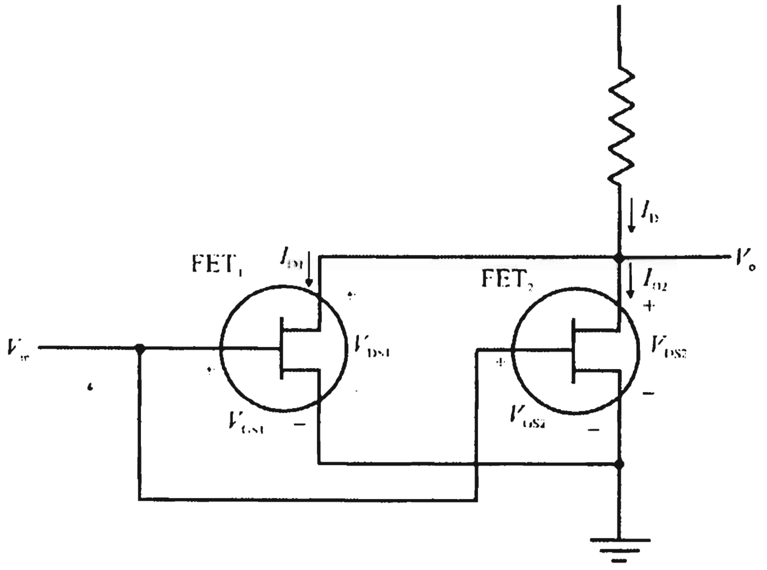


Figure C.16 If the dominant failure mode of the FETs is open drain to source, then a parallel connection such as this will allow the circuit to continue to operate in the event of a failure.

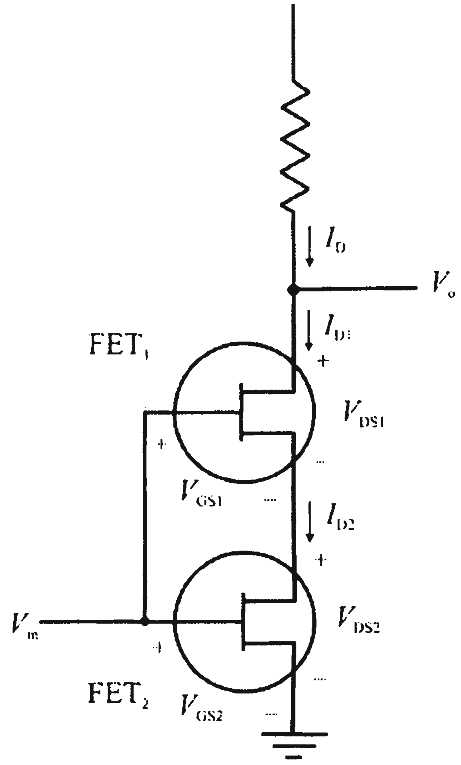


Figure C.17 Series configuration of FETs will protect against failures if the dominant failure mode is a short drain to source channel.

Analysis of the reliability of this circuit is the same as above. Likewise, if the dominant mode of failure of the transistor is shorting, the circuit shown in Figure C.17 will facilitate fault tolerance that can be shown as above. If the FET-dominant failure mode is not known, then the circuit shown in Figure C.18 can be implemented. In this case any one of the FETs can short or open and the circuit will continue to function correctly.

Extending these ideas to systems is straightforward except that the failure modes are not so easily dealt with. System components normally do not have such simple failure modes as open or short circuits that can be identified as such. Therefore, the failure modes do not normally enter into consideration in the same way.

Shown in Figure C.19 is a configuration of a subsystem that is critical to the operation of a system. In this system, three (or another odd multiple) of the critical subsystems are performing the same operations in time synchronization. The outputs of the three subsystems are continuously compared, and voting determines the response. If two or more of these outputs are the same, then that is assumed to be the correct subsystem response. If one of the outputs is in disagreement with the other two, then corrective actions are necessary to find out why this is the case. If it is determined that the subsystem has failed, then it would normally be removed,

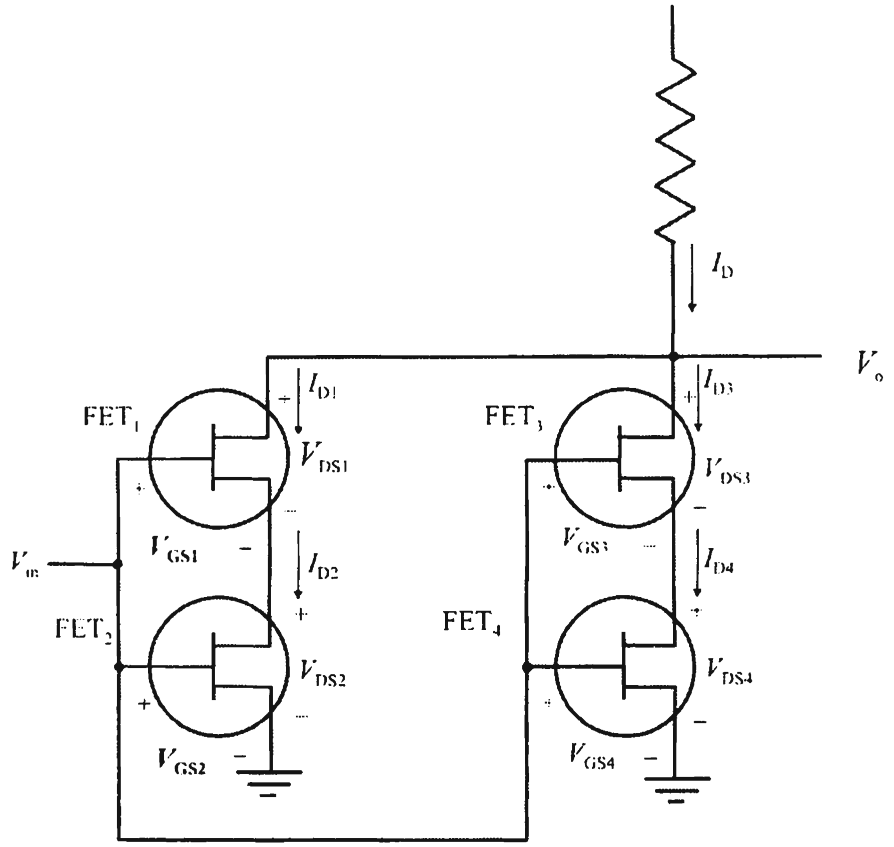


Figure C.18 A quadruple-redundant FET. Any one of the transistors can open or short and the circuit will continue to function.

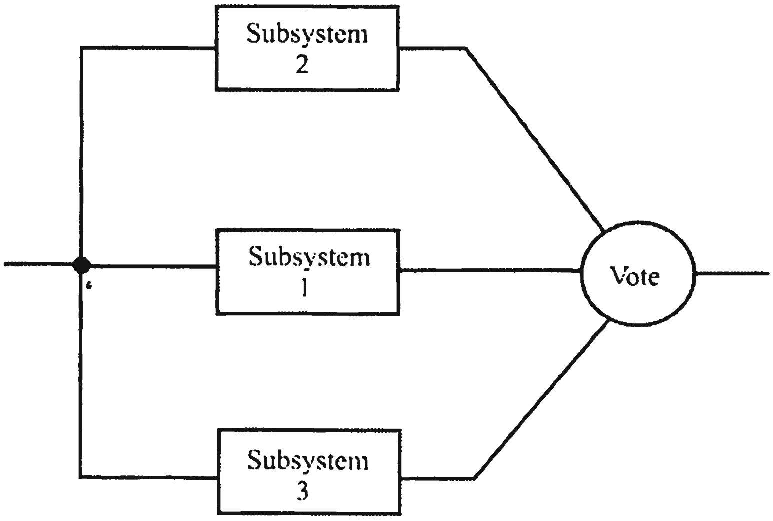


Figure C.19 A triple-redundant system. Each of the subsystems performs the same function and a vote of their outputs is taken. The largest vote is taken as the output.

with the resultant decrease in fault tolerance of the system. If all of the outputs disagree (for the first time), then it can be assumed that more than one of the subsystems has failed.

The reliability of the system shown in Figure C.19 is given by

$$R(t) = 1 - P(\text{failure}) \tag{C.8}$$

where failure will occur if all three subsystems fail or two out of the three fail. Therefore,

$$P(\text{failure}) = P(\text{all three fail}) + P(\text{two of the three fail}) \tag{C.9}$$

But

$$P(m\text{ out of }n\text{ fail}) = \binom{n}{m} p^{m}(1-p)^{n-m} \tag{C.10}$$

where $\binom{n}{m}$ is the binomial coefficient and p is the probability of any one of the subsystems failing, here assumed to be equal. Thus, since $R(t) = e^{-\lambda t}$, $p = 1 - e^{-\lambda t}$, and

$$\begin{aligned} P(\text{failure}) &= \frac{3!}{(3-3)!3!}(1-e^{-\lambda t})^{3}(1-e^{-\lambda t})^{3-3} \\ &\quad + \frac{3!}{(3-2)!2!}(1-e^{-\lambda t})^{2}(1-e^{-\lambda t})^{3-2} \end{aligned} \tag{C.11}$$

Carrying out the algebra yields

$$R(t) = 3e^{-2\lambda t} - 2e^{-3\lambda t} \tag{C.12}$$

This function is shown in Figure C.20 compared to the reliability of a single subsystem. Note that for short mission times, the triplicated system configuration provides higher reliability, but eventually falls below the single system reliability. This indicates that the fault tolerance is valid only for a limited amount of time.

The reason that the reliability eventually falls below the single subsystem reliability is that there are more components in the triplicated system. That is an unavoidable consequence of a more complicated system.

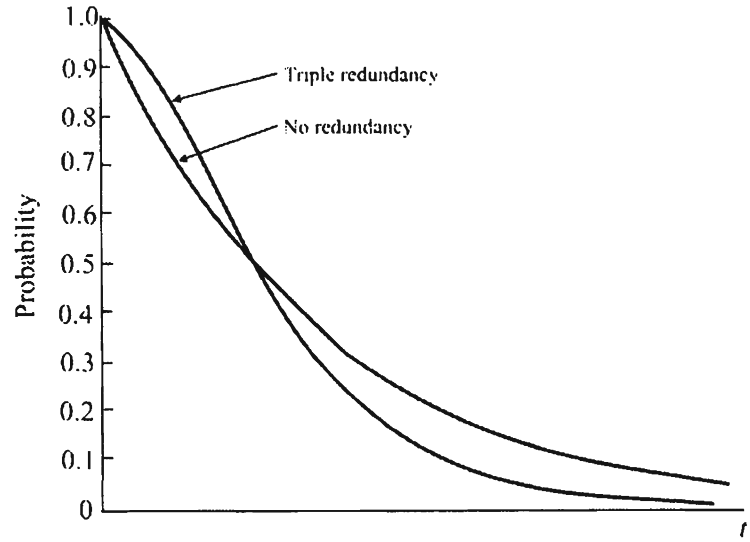


Figure C.20 Reliability function of the triple-redundant system shown in Figure C.19. Note that as time goes on, it becomes less reliable than any one of the subsystems alone.

C.3 Human-Factors Engineering

Human factors refers to the operation and the maintenance activities of a system. The differences between these two categories are discussed next.

C.3.1 Operations

It is always necessary to employ *human-factors engineering* (HFE) whenever a human operator is to use an EW system or a maintenance technician is to repair it. One or both of these situations is almost always true (the only exception that is obvious is space applications).

IEEE Standard 1023–1988 [4] outlines six HFE aspects that need to be considered when designing nuclear power stations, which are directly applicable to communication EW systems. These areas are listed as follows:

1. Tasks; ‹
2. Environment;
3. Equipment;
4. Personnel;

C. (Nuclear) Operations;

6. Documentation.

The task aspect includes the allocation of functions in the system to either machines or humans, which, of course, depends on whether the task can be automated or should be automated. It also involves the loading imposed on the operator of the system. The operator can be overloaded with tasks to be performed, or he or she can be under loaded; both conditions lead to underperformance. Also, sometimes machines can be more precise in their actions than humans can, and, therefore, if the task requires more precision than a human can deliver, it is a task that is a candidate for automation. The feedback to a human on the performance of a task is an important consideration.

In order to be useful the feedback should be timely and significant relative to how well the human performed the task. In addition, humans tend to be error-prone, and therefore any tasks allocated to a human must be tolerant of this error source. Lastly, the training of humans needs to be considered. It must be thorough and geared toward the level of education the expected system operators will have.

Humans can tolerate only certain variability in their physical environment. Factors such as temperature, humidity, and airflow are important considerations for the environment in which a human is to perform. Communication EW systems normally consist of extensive amounts of electronic equipment that generate heat. This heat source must be taken into account, along with other sources, such as Sun loading, to create an acceptable environment for human operators to occupy. Lighting must be carefully designed so that screen displays as well as visual indicators on the equipment can be easily read. Lighting should be adequate for the normal operation of the system. If computer displays are involved, the lighting should be on the dim side, allowing the screens to be viewed more clearly. Lights glaring on the screen should be avoided. The reasonable limits on viewing angles are shown in Figure C.21. These parameters should only be used when the normal limits shown above cannot be attained. These parameters refer to when the head is rotated to visualize displays. Acoustic noise should be minimized so that the

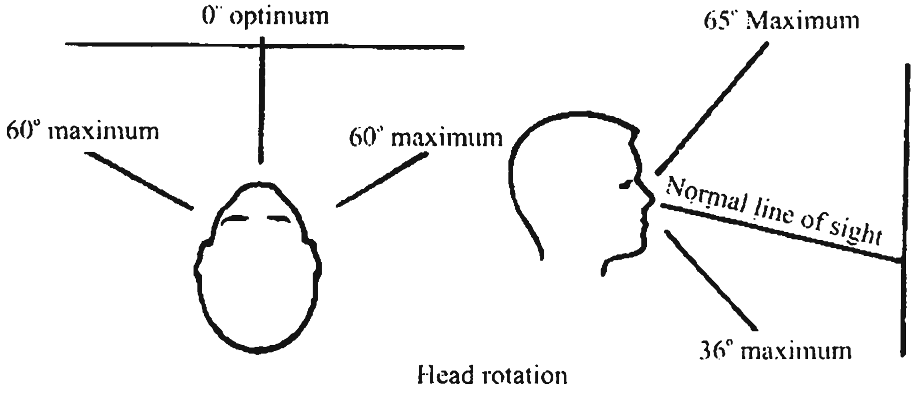
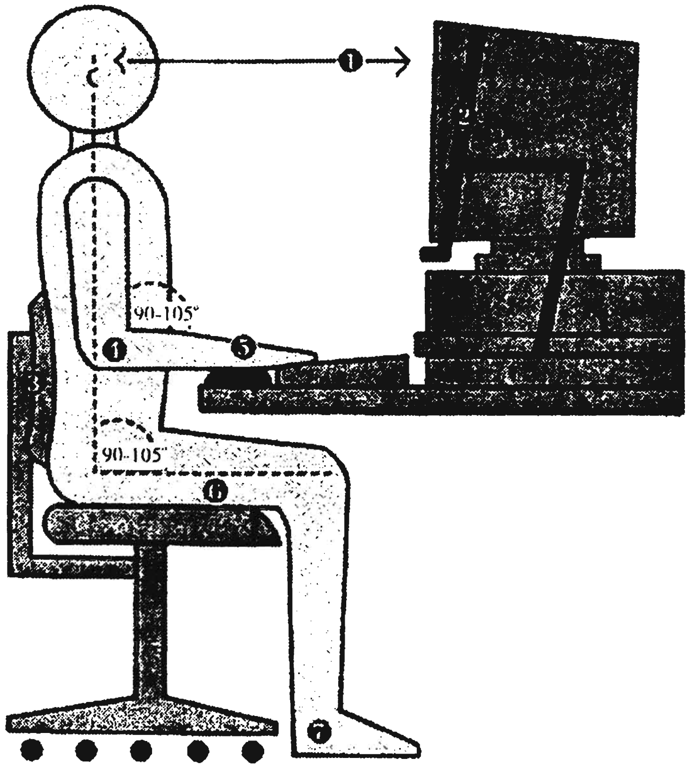


Figure C.21 Visual ranges for humans. (*Source:* [5]. © Association of Old Crows 1990. Reprinted with permission.)



1. Top of screen at eye level for bifocal wearers. Screen distance at arm's length (15-32").
2. Document holder adjustable to screen height.
3. Chair backrest provides firm lower back support. Chair back and seat easily adjustable for height and tilt by user.
4. Keyboard height promotes relaxed arms with forearms parallel to floor.
5. Wrists straight (neutral). Padded, movable wrist rest, same height as keyboard home row, if needed.
6. Thighs parallel to floor. Ample legroom under work surface.
7. Feet rest firmly on floor or foot rest.

Prepared by the Campus Occupational Health Program 1992

Figure C.22 Workstation layout. (*Source:* [6].)

operator is not distracted from performance of their normal duties. The size of the area in which the operator is to work needs to be adequate for the performance of the expected tasks.

Layout of equipment must be compatible with the tasks, so that timely and convenient execution of tasks can be performed. Equipment placement for tasks that the operator is to perform frequently should be placed as shown in Figure C.22, which represents a typical EW system operator console [6]. Specific concerns about equipment placement are illustrated in the figure. Operator positions are normally designed around ninety-fifth percentile humans, be they male or female. Displays must be located for easy viewing, especially if they are to be seen much. The most frequently used controls, such as keyboards, need to be within easy reach and at the right height. The best height for computer keyboards is between 26 and 3,0 inches, depending on the height of the operator. Therefore, this height must be adjustable. Use of color both for displays as well as warnings and indicators is useful. Green, yellow, and red lights display the obvious meaning at a glance. Finally, safety is always a significant consideration. Exposure to unsafe conditions should be minimized. The capabilities of the personnel expected to operate the system need to be taken into consideration. To a large extent, for military systems, this concern is addressed in a larger context when the operator personnel are trained in their *military occupational specialty* (MOS). There are limits, however, to human capabilities. If lights and indicators are too dim or poorly placed, eyestrain could result, and in the limit, the light may not be seen at all. The amount of information a human operator must remember is a consideration.

The final aspect considered in IEEE Standard 1023–1988 is documentation. The military services have set requirements for operation, maintenance, and training documentation. These are normally acquired when the system is first acquired, and updated as the system is updated. Needless to say, however, this documentation must be thorough and accurate. It must reflect all that an operator is expected to do under all conditions.

C.3.2 Maintenance

Frequently overlooked in the design of systems is the maintainability design and, in particular, the human factor aspect. How well a system can be maintained is directly related to how well a maintenance technician can repair it. "Well" in this case refers to the ability to determine what is wrong and to repair faulty equipment. A familiar example of this is the placement of oil filters in automobiles. A good design locates these filters where they are readily accessible—after all, it is known from the start that they will be replaced frequently, so why place them at difficult locations? A poor design, from a maintenance point of view, locates the filters at locations that are difficult to get to with the necessary wrench to loosen them.

It is not always possible to design systems so that all the possible maintenance issues are optimized. Some design requirements just do not allow it. Designing systems with maintainability in mind, however, is prudent and should always be considered.

C.3.3 Safety

Safety in operation as well as maintenance of a system is of prime importance. Safety issues can arise due to radiation effects, either intentional or unintentional. EA systems, for example, are designed to radiate high levels of energy and, depending on the frequency range, can cause injury to humans. Such a concern is exemplified by the operation of a high-power jammer at a frequency range that could potentially injure personnel. The levels of radiation that humans can tolerate for 0.1 hour according to ANSI Standard C9C.1-1982 are shown in Figure C.23 [7]. High voltages are almost always present in EW systems. Contact with these voltages can cause serious injury to both operators as well as maintenance personnel. Wherever high voltages are present, if possible, covers should be used. If not possible, then clear placards must be posted to minimize contact.

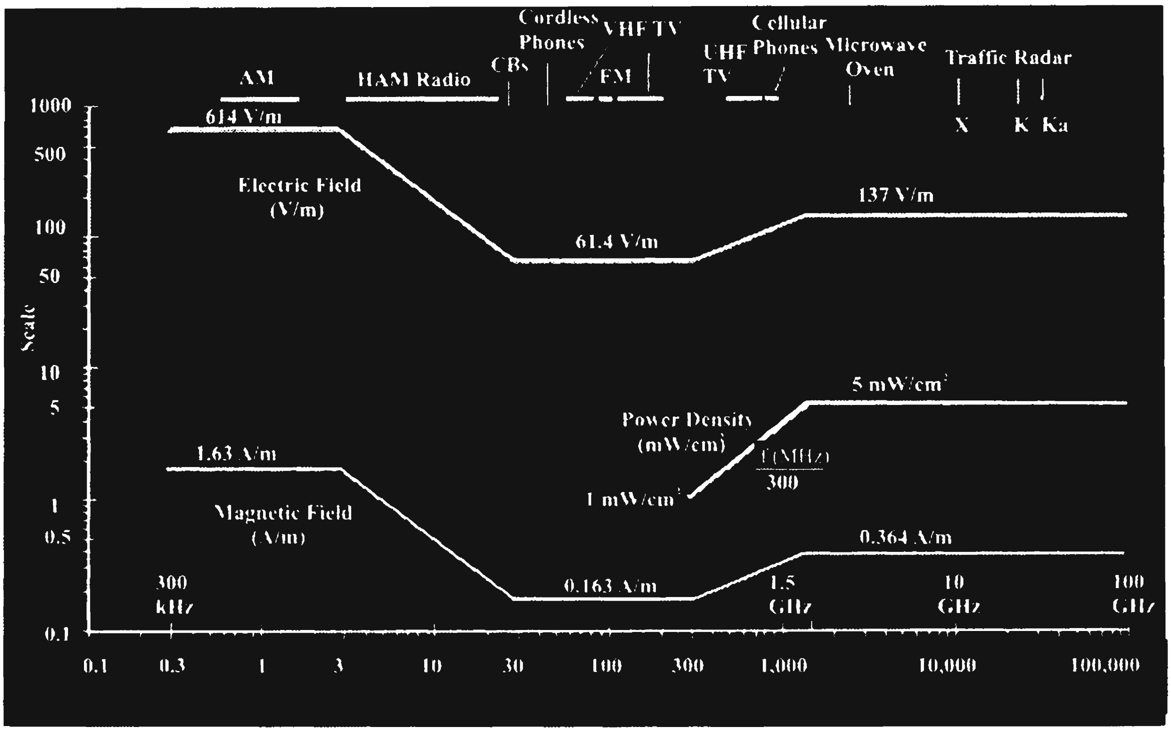


Figure C.23 Safety limits on the exposure of humans to radiation from ANSI Standard C9C.1-1982.

Some units are heavy, and require two or more people to handle them. Such cases would be clearly marked as such. When equipment is mounted in racks, heavier units should be placed lower in the rack if possible, precluding the requirement to hold heavy weights high in the air.

If the system allows it, two exits should always be provided. One of these would be the normal entryway, while the other could be an emergency exit in the ceiling, for example.

C.4 System Cost

Not much will be said here regarding the cost of communication EW systems, since the details of costs are entirely dependent on the details of the program to develop and deploy the system. A few comments are, however, general in nature.

The total cost of ownership of a system consists of two distinct parts: (1) the initial acquisition costs and (2) the sustainment costs. The first of these addresses what it costs to design and build the systems. It consists of the R&D phase as well as the quantity production. The second phase consists of the costs involved with keeping the systems operational once fielded. There are many cost elements to the second category, from training operators and maintainers to buying an inventory of spare parts.

In almost all cases, the sustainment costs are orders of magnitude larger than the initial acquisition costs. Therefore, it is important that during the design phase, adequate consideration be given to lowering the sustainment costs. One way of doing this, for example, is to use components or parts that already have been used in other systems so that the infrastructure support is already in place. Using the same part in multiple places reduces the cost of the inventory of spare parts.

C.5 Concluding Remarks

System engineering as a discipline applies engineering principles to the overall design of systems. A system in this case can be just about anything, but herein it generally refers to a complicated electronic ensemble of equipment. There is a process which, when followed, leads to an orderly life-cycle process of system design, construction, support, and ultimate disposal.

The environment in which a military system is used determines the testing required. In some cases benign environments apply where equipment is installed in office buildings and never moved, for example. The other extreme consists of muddy foxhole deployment. The latter requires more stringent design and testing so that the equipment will continue to operate when required.

Designing equipment to be reliable is critical. Complex systems that have many single points of failure are assured to having small MTBFs. Those functions that are necessary for a system to complete its critical mission modes must have redundancy built in. Reliability is designed into a system from the start, not applied late in the design phase.

References

[1] Kayton, M., "A Practitioner's View of System Engineering," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 33, No. 2, April 1997, pp. 579–586.

[2] IEEE Standard 1220-1998, IEEE Standard for Application and Management of the Systems Engineering Process, 1998.

[3] IEEE Standard 1156.2-1996, "Environmental Specifications for Computer Systems," 1996.

[4] IEEE Standard 1023-1988, "IEEE Guide for the Application of Human Factors Engineering to Systems, Equipment, and Facilities of Nuclear Power Generating Stations," December 12, 1988.

[5] "EW Engineer's Handbook," Association of Old Crows, Alexandria, VA, 1990.

[6] "Ergonomics and VDT Use," Library of Congress Collections Services, VDT Ergonomic Committee, 1991–92.

[7] Chignell, M., "Lecture 12: Standards and Plans," accessed December 2007, http://peach.mie.utoronto.ca/courses/mie240/html/lecture12.html.

List of Acronyms

A/D	analog-to-digital
ACK	acknowledgment
ADC	analog to digital converter
ADPCM	adaptive differential pulse code modulation
A_{eff}	effective area of antenna
AGC	automatic gain control
AGL	above ground level
AJ	antijam
ALE	automatic link establishment
ALR	average likelihood ratio
AM	amplitude modulation
AMPS	advanced mobile phone system
ANSI	American National Standards Institute
AO	area of operations
AOA	angle of arrival
AOI	area of interest
ARDIS	advanced radio data information service
ARQ	automatic repeat request
ASK	amplitude shift key
AWGN	additive white Gaussian noise
BBN	broadband noise
BCH	Bose, Chaudhuri, and Hocquenghem
BER	bit error rate
BFSK	binary frequency shift key
BJT	bipolar junction transistor
BPE	best point estimate
bps	bits per second
BPSK	binary phase shift key
BS	base station
BW	bandwidth

C2	command and control
C2W	command and control warfare
CAF	cross-ambiguity function
CB	citizen band
CDAA	circularly disposed antenna array
cdf	cumulative distribution function
CDMA	code division multiple access
CDPD	cellular digital packet data
CEOI	communications electronics operating instructions
CEP	circular error probable
CFAR	constant false alarm rate
CNR	carrier-to-noise ratio
COP	common operational picture
CRC	cyclic redundancy check
CSMA	carrier sensed multiple access
CW	continuous wave
CWT	continuous wavelet transform

DAMA	demand assignment multiple access
DAC	digital to analog converter
dB	decibel
dBc	decibel relative to carrier
dB$_{\text{kTB}}$	decibels relative to kTB (thermal noise floor)
dBm	decibel relative to 1 milliwatt
DBK	dominant battlespace knowledge
DBPSK	differential binary phase shift key
DCPM	differential pulse code modulation
DCT	discrete cosine transform
DD	differential Doppler
DES	data encryption standard
DF	direction finding
DSB	double sideband
DMT	discrete multitone
DoD	Department of Defense
DPCM	differential pulse code modulation
DPSK	differential phase shift key
DSA	digital signatures algorithm
DSS	digital signature standard

DSSS	direct-sequence spread spectrum
DWT	discrete wavelet transform
EA	electronic attack
EEP	elliptical error probable
EES	escrow encryption standard
EHF	extra-high frequency
EM	electromagnetic
EMC	electromagnetic compatibility
EMCON	emission control
EMI	electromagnetic interference
EO	electro-optical
EOB	electronic order of battle
EP	electronic protect
ERP	effective radiated power
ES	electronic support
ESM	electronic support measures
EW	electronic warfare
FAIT	first article integration and test
FCC	Federal Communication Commission
FCS	frame check sequence
FDD	frequency division duplexing
FDDI	fiber distributed data interface
FDM	frequency division multiplexing
FDMA	frequency division multiple access
FDOA	frequency difference of arrival
FEC	forward error correction
FET	field effect transistor
FFH	fast frequency hopping
FFT	fast Fourier transform
FH	frequency-hopping
FHSS	frequency-hopping spread spectrum
FLOT	forward line of own troops
FM	frequency modulation
FSK	frequency shift key
GDOP	geometric dilution of precision
GHz	gigahertz (10^9 hertz)
GMSK	Gaussian minimum shift key
GPS	global positioning system

GSM	global system for mobile communication
HF	high frequency
HFE	human-factors engineering
HOS	higher order spectra
HPA	high-power amplifier
Hz	hertz (cycles per second)
HUMINT	human intelligence
IBW	instantaneous bandwidth
IEEE	Institute of Electrical and Electronic Engineers
IF	intermediate frequency
IFBW	intermediate-frequency bandwidth
IFFT	inverse fast Fourier transform
i.i.d.	independent and identically distributed
INTELSAT	international telecommunication satellites
IN	impulsive noise
IR	infrared
ISI	intersymbol interference
ISDN	integrated services digital network
ISI	intersymbol interference
ISM	instrumentation, scientific, and measurement
ITU	International Telecommunication Union
IW	information warfare
J/S	jammer power-to-signal power ratio
JPEG	Joint Pictures Expert Group
JSR	jammer to signal power ratio
kbps	kilobits per second (10^3 bps)
km	kilometers
kHz	kilohertz (10^3 hertz)
KLT	Karhunen-Loeve transform
LEAF	law enforcement access field
LF	low frequency
LFSR	linear feedback shift register
LOB	line of bearing
LOP	line of position
LOS	line of sight

LPD	low probability of detection
LPE	low probability of exploitation
LPI	low probability of intercept
LQA	link quality analysis
LSB	least significant bit
LSB	lower sideband
LUF	lowest usable frequency
MAP	maximum a posteriori probability
Mbps	megabits per second (10^6 bps)
MEM	microelectrical-mechanical
MF	medium frequency
MFSK	multiple frequency shift key
MHz	megahertz (10^6 hertz)
Mil Std	military standard
MIT	Massachusetts Institute of Technology
MLR	maximum likelihood ratio
MOE	measure of effectiveness
MOS	military occupational specialty
MPEG	Motion Pictures Expert Group
MPSK	multiple phase shift key
ms	millisecond (10^{-3} seconds)
MS	mobile station
MSE	mean square error
MSK	minimum shift key
MTBF	mean time between failures
MTI	moving target indicator
MTTR	mean time to repair
MUF	maximum usable frequency
MUSIC	multiple-signal classification
NACK	negative acknowledgment
NCBFSK	noncoherent binary frequency shift key
NCFSK	noncoherent frequency shift key
NCS	net control station
NF	noise figure
NIST	National Institute of Standards and Technology
NSA	National Security Agency
NVIS	near-vertical incidence skywave

O&S	operations and support
OFDM	offset frequency division multiplexing
OFDMA	orthogonal frequency division multiple access
OOK	on-off key
OOTW	operations other than war
OPSEC	operational security
OQPSK	offset quadrature phase shift key
OTM	on the move
PA	power amplifier
PAM	pulse amplitude modulation
PBC	phase-based classifier
PB-FBC	partial band filter-bank combiner
PBN	partial band noise
PC	personal computer
PCM	pulse code modulation
PCS	personal communication system
pdf	probability density function
PF	position fix
PGP	pretty good privacy
PHP	personal handy phone
PLL	phase-locked loop
PM	phase modulation
PN	pseudo-noise
POI	probability of intercept
PPM	pulse position modulation
psd	power spectral density
PSK	phase shift key
PSP	per-survivor processing
PSTN	public switched telephone network
PSYOPS	psychological operations
PTT	push-to-talk
QAM	quadrature amplitude modulation
QPSK	quadrature phase shift key
R&D	research and development
RF	radio frequency
RFI	radio frequency interference
RLOS	radio line of sight
RMS	root mean square

ROC	receiver operating characteristic
RS	Reed Soloman
r.v.	random variable

S/H	sample and hold
SAM	surface to air missile
SAR	synthetic aperture radar
SC	suppressed carrier
SEP	system engineering process
SFH	slow frequency hopping
SHA	secure hash algorithm
SHF	superhigh frequency
SIGINT	signals intelligence
SINCGARS	single channel ground and air radio system
SJR	signal-to-jam ratio
SLQ	square law classifier
SNR	signal-to-noise ratio
SOI	signal of interest
SSB	single sideband
STFT	short-term Fourier transform
SVD	singular-value decomposition

TBW	total bandwidth
TCC	Turbo-coded channel
TCM	trellis-coded modulation
TDD	time division duplexing
TDMA	time division multiple access
TDOA	time difference of arrival
TOA	time of arrival
TOC	tactical operations center
TSS	telecommunications standardization sector
TTIP	transparent tone in band
TV	television
TWT	traveling wave tube

UAS	unmanned aerial system
UHF	ultrahigh frequency
USB	upper sideband
V	volt
VCO	voltage-controlled oscillator
VHF	very high frequency

VLF	very low frequency
VME	versa-module Europe
WADF	wide aperture direction finder
WGN	wideband Gaussian noise
W	watts
WW	World War

About the Author

Richard A. Poisel matriculated with a B.S. in electrical engineering from the Milwaukee School of Engineering in 1969 and received an M.S. in the same discipline from Purdue University in 1971. He spent three years in military service from 1971 to 1973. After this service he attended the University of Wisconsin, where he received a Ph.D. in electrical and computer engineering in 1977. From 1977 to 2004 he was with the same government organization, which has had several different names and is currently known as the U.S. Army Research, Development, and Engineering Command, Intelligence and Information Warfare Laboratory. Dr. Poisel was the director of the laboratory from 1997 to 1999. During the 1993–1994 academic year, Dr. Poisel attended the MIT Sloan School of Management as a Sloan Fellow, and received an M.B.A. He is currently employed by Raytheon Missile Systems in Tucson, Arizona.

Index

The Artech House Intelligence and Information Operations Series

Concepts, Models, and Tools for Information Fusion, Éloi Bossé, Jean Roy, and Steve Wark

Electronic Intelligence: The Analysis of Radar Signals, Second Edition, Richard G. Wiley

Electronic Warfare for the Digitized Battlefield, Michael R. Frater and Michael Ryan

Electronic Warfare in the Information Age, D. Curtis Schleher

Electronic Warfare Target Location Methods Second Edition, Richard A. Poisel

EW 101: A First Course in Electronic Warfare, David Adamy

Homeland Security Technology Challenges: From Sensing and Encrypting to Mining and Modeling, Giorgio Franceschetti and Marina Grossi, editors

Information Warfare and Organizational Decision-Making, Alexander Kott, editor

Information Warfare Principles and Operations, Edward Waltz

Introduction to Communication Electronic Warfare Systems, Second Edition, Richard A. Poisel

Knowledge Management in the Intelligence Enterprise, Edward Waltz

Mathematical Techniques in Multisensor Data Fusion, Second Edition, David L. Hall and Sonya A. H. McMullen

Modern Communications Jamming Principles and Techniques, Richard A. Poisel

Principles of Data Fusion Automation, Richard T. Antony

Strategem: Deception and Surprise in War, Barton Whaley

Statistical Multisource-Multitarget Information Fusion Ronald P. S. Mahler

Tactical Communications for the Digitized Battlefield, Michael Ryan and Michael R. Frater

Target Acquisition in Communication Electronic Warfare Systems, Richard A. Poisel

For further information on these and other Artech House titles, including previously considered out-of-print books now available through our In-Print-Forever® (IPF®) program, contact:

Artech House
685 Canton Street
Norwood, MA 02062
Phone: 781-769-9750
Fax: 781-769-6334
e-mail: artech@artechhouse.com

Artech House
46 Gillingham Street
London SW1V 1AH UK
Phone: +44 (0)20-7596-8750
Fax: +44 (0)20-7630-0166
e-mail: artech-uk@artechhouse.com

Find us on the World Wide Web at: www.artechhouse.com

Tactical Communications for the Digitized Battlefield, Michael Ryan and Michael R. Frater

Target Acquisition in Communication Electronic Warfare Systems, Richard A. Poisel

For further information on these and other Artech House titles, including previously considered out-of-print books now available through our In-Print-Forever® (IPF®) program, contact:

Artech House
685 Canton Street
Norwood, MA 02062
Phone: 781-769-9750
Fax: 781-769-6334
e-mail: artech@artechhouse.com

Artech House
46 Gillingham Street
London SW1V 1AH UK
Phone: +44 (0)20-7596-8750
Fax: +44 (0)20-7630-0166
e-mail: artech-uk@artechhouse.com

Find us on the World Wide Web at: www.artechhouse.com